

# An Evolutionary Model of Cooperation, Fairness and Altruistic Punishment in Public Good Games

Moritz Hetzer<sup>1\*</sup>, Didier Sornette<sup>1,2</sup>

**1** Chair of Entrepreneurial Risks, Department of Management, Technology and Economics, ETH-Zurich, Zurich, Switzerland, **2** Swiss Finance Institute, c/o University of Geneva, Geneva, Switzerland

## Abstract

We identify and explain the mechanisms that account for the emergence of fairness preferences and altruistic punishment in voluntary contribution mechanisms by combining an evolutionary perspective together with an expected utility model. We aim at filling a gap between the literature on the theory of evolution applied to cooperation and punishment, and the empirical findings from experimental economics. The approach is motivated by previous findings on other-regarding behavior, the co-evolution of culture, genes and social norms, as well as bounded rationality. Our first result reveals the emergence of two distinct evolutionary regimes that force agents to converge either to a defection state or to a state of coordination, depending on the predominant set of self- or other-regarding preferences. Our second result indicates that subjects in laboratory experiments of public goods games with punishment coordinate and punish defectors as a result of an aversion against disadvantageous inequitable outcomes. Our third finding identifies disadvantageous inequity aversion as evolutionary dominant and stable in a heterogeneous population of agents endowed initially only with purely self-regarding preferences. We validate our model using previously obtained results from three independently conducted experiments of public goods games with punishment.

**Citation:** Hetzer M, Sornette D (2013) An Evolutionary Model of Cooperation, Fairness and Altruistic Punishment in Public Good Games. PLoS ONE 8(11): e77041. doi:10.1371/journal.pone.0077041

**Editor:** Yamir Moreno, University of Zaragoza, Spain

**Received:** May 12, 2013; **Accepted:** September 5, 2013; **Published:** November 19, 2013

**Copyright:** © 2013 Hetzer, Sornette. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** The authors have no support or funding to report.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: moritz.hetzer@gmail.com

## Introduction

Why do we maintain moral attitudes, display other-regarding behavior, have a distaste for unfairness, act prosocially and, at times, even behave altruistically towards others? How is this behavior compatible with the predominant theories of rational choice, selfish utility maximization and, in particular, with Darwin's principle of the survival of the fittest? This article presents an evolutionary utility framework of fairness, altruistic punishment and cooperation. It develops quantitative arguments supporting the hypothesis that the key to understanding the ostensibly mysterious patterns of human behavior is deeply rooted in our evolutionary history.

The analysis of our expected utility model, in combination with the underlying evolutionary dynamics, allows us to identify and explain the origin and the emergence of other-regarding preferences and, ultimately, enables us to quantitatively explain the degree of altruistic punishment that is observed in lab experiments. As a result, our approach complements and extends other utility frameworks, e.g. the Fehr Schmidt model [1], Bolton/Ockenfels [2] and Rabin [3], as well as existing evolutionary models linked to economics [4–10], by combining both perspectives in order to explain prosocial behavior. Unlike other approaches, our model does not assume *ex ante* the existence of other-regarding preferences, but instead demonstrates their co-evolutionary emergence along with the emergence of altruistic punishment behavior. The design of our model is inspired by previous findings about the co-evolution of culture, norms and

genes, the effect of other-regarding behavior as well as bounded rationality. We motivate our model by the psychological predisposition of individuals to maximize their expected utility together with subliminal disposition to follow social norms [11–15]. Both mechanisms are closely related in the process of gene-culture co-evolution.

The following section 0 describes the model in detail and explains the interplay of agents who maximize their expected utility under the effects of natural selection and competitive evolutionary dynamics. Then, section 0.0 presents empirical tests of the theory. Section 0.0 establishes the evolutionary dominance of the specific other-regarding preference in the form of disadvantageous inequity aversion. Section 0.0 concludes.

## The Model

### 1 General framework

We take an evolutionary approach as a starting point to construct our model. The fitness of an agent is considered to be equivalent to her realized cumulative payoff relative to that of other agents, i.e. to the monetary units (MU) that the agent gains over time relative to the average over all other agents. Each agent  $i$  is characterized by one or multiple traits. The traits of an agent determine her behavior and correspond to a pure strategy denoted by  $s_i$ . Traits are passed on as fitness weighted values to the offspring in the process of evolutionary reproduction. The population is determined by the set of pure strategies  $S$ . In an evolutionary competitive environment, agents are subject to

natural selection which affects their viability and fertility. While viability selection accounts for removing poor performing agents from the population, fertility selection enables more successful agents to spread and to promote their genetic and cultural heritage in the population. This process corresponds to the standard evolutionary challenge of survival and reproduction. Following the Darwinian principle of the survival of the fittest, both selection mechanisms are defined relative to the environment of an agent. This means that the fitness of an agent is determined relative to the performance of the remaining population that she is exposed to and interacts with. In an evolutionary environment, the success of an agent and of her strategies defines the fitness of the agent and thus determines the proportional change of the strategies (traits) in the population over time.

The set of strategies  $\mathcal{S}$  that characterizes a population of agents is specified by a probability measure  $P^t$  that quantifies the frequencies of the single strategies  $s_i \in \mathcal{S}$  in the population at time  $t$ . In the two player case, the payoff of an agent who plays a pure strategy  $s \in \mathcal{S}$  against another agent who plays the pure strategy  $\hat{s}$  is denoted by  $f(s, \hat{s})$ . Both,  $s$  and  $\hat{s}$  are defined in the continuous strategy space  $\mathcal{S} \subset \mathbb{R}^x$  ( $x$  depends on the game; in the public good game studied here,  $x=2$  corresponds to the decision on the contribution level  $m$  and on the chosen propensity  $k$  to punish for each agent). For the  $n$ -player case, the average payoff of an agent who plays a strategy  $s$  at time  $t$  against a population characterized by the probability measure  $P^t$  over the strategy space  $\mathcal{S}$  is defined by

$$E(s, P^t) = \int_{\mathcal{S}} f(s, \hat{s}) P^t(d\hat{s}). \quad (1)$$

The total average payoff of the entire population at time  $t$  is defined by

$$E(P^t, P^t) = \int_{\mathcal{S}} \int_{\mathcal{S}} f(s, \hat{s}) P^t(d\hat{s}) P^t(ds). \quad (2)$$

The success of a strategy  $s$  is given by the difference of equations (1) and (2) as shown e.g. in [16–18]:

$$\begin{aligned} \Phi(s, P^t) &= E(s, P^t) - E(P^t, P^t) \\ &= \int_{\mathcal{S}} f(s, \hat{s}) P^t(d\hat{s}) - \int_{\mathcal{S}} \int_{\mathcal{S}} f(s, \hat{s}) P^t(d\hat{s}) P^t(ds) \end{aligned} \quad (3)$$

The dynamics of the frequency of a specific strategy  $s$  in the population is defined by the ordinary differential equation

$$dP^t(ds)dt = \Phi(s, \hat{s}) \cdot P^t(ds). \quad (4)$$

The function  $f(s, \hat{s})$  in equations (1–3) reflects the underlying payoff structure of the analyzed game. In the context of this paper,  $f(s, \hat{s})$  represents the payoff function of a public goods game with punishment.

## 2 The public good games with punishment

In the following, we model the behavior of agents playing a standard one-shot-interaction public goods game with punishment as presented in [19–21]. Agents are pooled in groups of size  $n$ . Each agent  $i$  is characterized by a strategy  $\hat{s}_i = [m_i, k_i]$  that is defined by two traits. The first trait  $m_i$  corresponds to the amount

of MUs an agent contributes to the common group project (the public good) and thus reflects the agent's willingness to cooperate. The second trait  $k_i$  reflects the agent's propensity to punish defectors in the group.

The general procedure of the public goods game with punishment is as follows. In the first stage of the game, agent  $i$  contributes  $m_i$  monetary units (MUs) to a common public good which yields a return of  $g$  MUs per invested MU. The return from the public good is equally redistributed among the  $n$  group members. Agents then learn about the contributions of the other group members. In a second stage, they are provided with the opportunity to punish other group members. Punishment comes in the form of additional costs for both the punisher as well as the punished agent: for each MU spent by the punisher, the return that the punished agent obtained from the public goods game is reduced by  $r$  MUs. Given the one-shot-interaction characteristic of the game, punishment does not result in a direct or indirect material benefit and is often considered in the literature to be an altruistic act.

## 3 Modeling assumptions

We make the following assumptions about the behavior of agents and the evolutionary environment:

- Agents are assumed to be self-interested and to act rationally given their available information and computational capabilities [22–25]. In particular, agents are involved in one-shot interactions only and have no ex-ante information about the others' actions at the time they take their decisions. Thus, other agents' past actions (history) will not affect the agent's current or future decisions. This corresponds to the so-called stranger treatment in the experiments that we analyze below. This can also be called the strong mixing regime with rapid memory loss. A perhaps more convincing interpretation of this framework is in terms of a coarse-grained description of the multiple interactions and feedbacks between agents within groups that act over time scale of generations. In this interpretation, the unit time step is roughly commensurate with the agent lifetime. Then, other agents' past actions and history occur essentially within each time step but not beyond, justifying the model assumption.
- Agent  $i$  is assumed to punish agent  $j$  according to a function that is linearly increasing with the negative deviation between  $j$ 's and  $i$ 's contributions. Specifically, if  $m_j - m_i < 0$ , agent  $i$  punishes agent  $j$  by spending  $k \cdot (m_i - m_j)$  MUs, while  $j$  suffers a loss of  $r \cdot k \cdot (m_i - m_j)$  MUs. We assume this linear dependency because it can frequently be observed in experiments conducted in the western cultural area [19–21, 26, 27]. Figure 1 illustrates this behavioral pattern for data obtained in three public goods games [19–21].
- The factor  $k$  describes the agents' propensity to punish negative deviators.  $k$  is assumed to be a common trait or a norm that is shared by all agents within a homogeneous population. It reflects the subjects' genetically and culturally encoded behavior to react to actions that are e.g. perceived as being unfair. The interplay of punishment and evolutionary dynamics (adaptation, selection, cross-over and mutation) over hundreds of thousands of years caused the convergence of a previously diverse set of behavioral patterns. This process ultimately led to a common set of behavioral traits which are shared among directly- or indirectly-related and -interacting individuals, e.g. groups originating from the same cultural area. Vice-versa, the prevalent set of behavioral traits determined the anticipated expectations about the behavior

of individuals from the same cultural and genetic background. Punishment thus provided the basis for the emergence and manifestation of traits and (social) norms, while simultaneously punishment itself got frequently established as a common trait and norm. In conclusion, humans and our ancestors have converged and evolved to this common norm-enforcing feedback mechanism over hundreds of thousands of years as a result of gene-culture co-evolutionary processes [28–30]. The subjects’ psychological predispositions to render these encoded norms effective ultimately results in the focal action that is observed as a direct and immediate harm towards negative deviators or it acts as a hidden deterrence [11]. Today, lab experiments and field studies such as those of [19–21,31,32] allow one to sample and observe the statistically stationary characteristics of the common propensity to punish  $k_i$  from subjects originating from a similar cultural background.

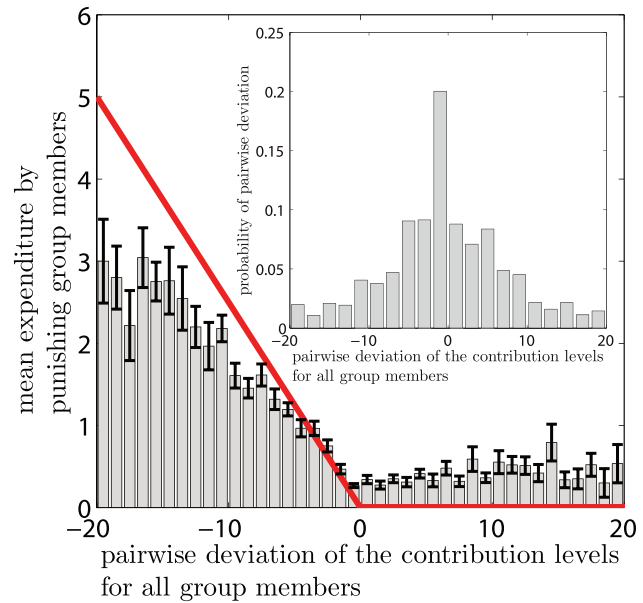
- The population of agents is subject to evolutionary dynamics in the form of selection, cross-over and mutation. These three mechanisms affect the viability and fertility of an agent. Viability selection induces a minimal survival condition in the form of a fixed lower value of consumption  $c_{fix}$ . This value reflects the basic requirements of an agent, i.e. it defines a lower limit that an agent needs to consume per unit of time in order to survive.  $c_{fix}$  thus repeatedly absorbs a fraction of the agents’ fitness value. Fertility selection accounts for the selection of successful genotypes, i.e. strategies, as opposed to unsuccessful ones. Agents can spread their strategies in the population proportionally to their fitness, e.g. by producing more offsprings. The relative change of the frequency of a trait, i.e. a strategy, is determined by the average success of that trait with respect to the average success of the remaining traits in the population. Cross-over, i.e. the reproduction by mating of two or more agents in the population towards those strategies that are carried by more successful agents. In contrast, mutation induces an additional heterogeneity to the agents’ strategy pool and allows the population to explore further the potential strategy space. This ensures that a population of agents is always heterogeneous with respect to the strategies, i.e.  $\text{Variance}(m_i) > 0$  and  $\text{Variance}(k_i) > 0$ .

#### 4 Utility formulation of the public goods game model

We first formulate a fitness model assuming complete information. The profit and loss (P&L) of an agent  $i$  that plays a public goods game with punishment is determined by the payoff from the game minus the costs of punishing and being punished and minus the contributed effort:

$$\begin{aligned}
 f_i(m_1, \dots, m_n) = & -m_i + \frac{g}{n} \cdot (m_i + \sum_{j \neq i} m_j) \\
 & -k \cdot r \cdot \sum_{j \neq i} \max(m_j - m_i, 0) \\
 & -k \cdot \sum_{j \neq i} \max(m_i - m_j, 0)
 \end{aligned} \tag{5}$$

The first term in the right hand side of equation (5), i.e.  $m_i$ , corresponds to the contribution of agent  $i$  to the public good. The second term represents the return from the public good. The third and fourth terms quantify the costs of being punished by others and punishing others, respectively. The number of agents in the group is denoted by  $n$ , the return from the public good is  $g$  per



**Figure 1. Mean expenditure of a given punishing member as a function of the deviation between her contribution and that of the punished member, for all pairs of subjects within a group, as reported empirically [19–21].** The error bars indicate the standard error around the mean. The straight line crossing zero with a slope of  $-k$  shows the average decision rule for punishment. The anomalous punishment of cooperators, corresponding to the positive range along the horizontal axis, is neglected in our model. The inset shows the relative frequency of the pairwise deviations. doi:10.1371/journal.pone.0077041.g001

invested MU, and  $r$  corresponds to the punishment efficiency factor.

Analogously, the P&L of the remaining agents  $j \neq i$  can be written as

$$\begin{aligned}
 f_j(m_1, \dots, m_n) = & -m_j + \frac{g}{n} \cdot (m_i + \sum_{j \neq i} m_j) \\
 & -k \cdot r \cdot \sum_{j \neq i} \max(m_j - m_i, 0) \\
 & -k \cdot \sum_{j \neq i} \max(m_j - m_i, 0).
 \end{aligned} \tag{6}$$

By writing the relative fitness of an agent according to equation (3) in the form of an evolutionary measure of success, we obtain the fitness of agent  $i$  as the sum of the experienced payoff differences between the own monetary payoff  $f_i$  and the monetary payoff of the remaining individual group members  $f_j$ :

$$u_i(f_1, \dots, f_n) = \sum_{j=1, \dots, n, j \neq i} (f_i - f_j) \tag{7}$$

The relative fitness of an agent is thus obtained by putting the payoff of agent  $i$  in relation to the payoff of the remaining population. This form of the fitness describes a population of agents that is exposed to evolutionary dynamics as presented in section 0.1: Positive values of  $u_i(f_1, \dots, f_n)$  are desirable, because they are associated with a higher fertility and a lower mortality.

Negative values of  $u_i(\dots)$  should be avoided in order to prevent the evolutionary extinction of the own traits.

In our attempt to combine preferences, expected utility and evolution, we should be careful and clarify what is meant by fitness and by utility. It is customary to distinguish fitness from utility, the former being proportional to material payoffs as defined above (eventually controlling the number and quality of offsprings), while utility refers to preferences. In full generality, it may be that preferences differ from payoffs, as when utility depends on the presence of salient unchosen alternatives that do not have any impact on the payoffs [9]. This distinction may lead to a fruitful strategy to understand how evolution processes may lead to apparent dysfunctional preferences [8] and contribute to explain the nature of our utility functions [9]. However, since our purpose is limited to provide a generalization of utility approaches of [1], [2] and [3] to explain the experimental results of three independently conducted experiments of public goods games with punishment, we shall neglect the difference between fitness and preference. For our purpose, agents in our game as well as in the experiments are not exposed to salient unchosen alternatives, suggesting that we can neglect the distinction between fitness and utility within our framework. The issue of fitness versus utility falls within the bigger question of the so-called ‘‘adaptationist program’’, which has dominated evolutionary thinking, and its possible caveats have been insightfully presented by [33].

Henceforth, by substituting equations (5) and (6) into equation (7), we obtain what can be termed the evolutionary utility of an agent, given by the two-term function shown in equation (8) below. The first term of (8) corresponds to agent  $i$ 's utility gained from the payoff of the public goods game with punishment. The second term of equation (8) represents the payoff for each of the  $n - 1$  opponents indexed by  $j$ . The total utility for agent  $i$  is defined by the sum of the differences between all combinations of  $f_i(m_1, \dots, m_n)$  and  $f_j(m_1, \dots, m_n)$  with  $j \neq i$ :

$$u_i(f_1, \dots, f_n) = \sum_{j=1, \dots, n, j \neq i} (f_i(m_1, \dots, m_n) - f_j(m_1, \dots, m_n)) \quad (8)$$

Consistent with utility theory (even in the presence of bounded rationality) and the underlying evolutionary dynamics, we assume that the agents seek to maximize their utility [22]. Obviously, the maximum of the utility function (8) can only be calculated in the hypothetical case of complete information about the others' contributions. However, information about the individual contributions  $m = (m_1, \dots, m_j)$  is not available ex ante, because agents decide about their contributions simultaneously. It follows that agents are required to make assumptions, i.e. to form their first-order beliefs, about the others' contributions. We model this by transforming the utility model in equation (8) into a subjective expected utility model.

Therefore, we introduce the subjective probability measure  $P_i(m_j)$  that represents agent  $i$ 's (first-order) belief about the contributions of the other agents.  $P_i(m_j)$  quantifies the likelihood as perceived by agent  $i$  that another agent  $j$  will contribute  $m_j$  MUs. In the one-shot game version studied here, all agents  $j \neq i$  are indistinguishable from the point of view of an agent  $i$ , i.e., agent  $i$  has no information on any preference, trait or specific characteristics of the other agents. Using  $P_i(m_j)$ , agent  $i$  can form her expectation [34] about the average of the other agents' contributions:

$$E_i[m_j] = \int m_j \cdot P_i(m_j) dm_j. \quad (9)$$

Similarly to the propensity  $k$  to punish,  $E_i[m_j]$  can be interpreted as the expected norm-conforming behavior of the population that has co-evolved, learned and internalized across time in a population of interacting agents.

The utility model defined in equation (8) is transformed into an expected utility model using the subjective expectations  $E_i[m_j]$ . Rewriting  $f_i(m_1, \dots, m_n)$  and  $f_j(m_1, \dots, m_n)$  by replacing each value  $m_j \in [m_1, \dots, m_{i-1}, m_{i+1}, \dots, m_n]$  with agent  $i$ 's subjective expectation  $E_i[m_j]$  on  $m_j$  gives the following equations:

$$\begin{aligned} E_i[f_i(m_i)] &= -m_i + \frac{g}{n} \cdot m_i \\ &+ \frac{g}{n} \cdot (n-1) \cdot \int_0^\infty m_j \cdot P_i(m_j) dm_j \\ &- (n-1) \cdot k \cdot r \cdot \int_{m_i}^\infty (m_j - m_i) \cdot P_i(m_j) dm_j \\ &- (n-1) \cdot k \cdot \int_0^{m_i} (m_i - m_j) \cdot P_i(m_j) dm_j \end{aligned} \quad (10)$$

$$\begin{aligned} E_i[f_j(m_i)] &= - \int_0^\infty m_j \cdot P_i(m_j) dm_j + \frac{g}{n} \cdot m_i \\ &+ \frac{g}{n} \cdot (n-1) \cdot \int_0^\infty m_j \cdot P_i(m_j) dm_j \\ &- k \cdot r \int_0^{m_i} (m_i - m_j) P_i(m_j) dm_j \\ &- k \cdot \int_{m_i}^\infty (m_j - m_i) P_i(m_j) dm_j \end{aligned} \quad (11)$$

Note that, in the formation of the expectation by agent  $i$  of the others' utility functions, agent  $i$ 's own contribution  $m_i$  is obviously known to her, hence the term  $\frac{g}{n} \cdot m_i$  appears without averaging.

As in the case of complete information, agents seek to maximize their relative fitness, i.e. the sum of the differences between their own P&L,  $f_i(m_1, \dots, m_n)$ , and the others' P&L. Putting all this together, we obtain the evolutionary expected utility function  $u_i(E_i[f_i(m_i)], E_i[f_j(m_i)])$  of agent  $i$  as shown in equation (12).

$$u_i(E_i[f_i(m_i)], E_i[f_j(m_i)]) = (n-1) \cdot (E_i[f_i(m_i)] - E_i[f_j(m_i)]) \quad (12)$$

We start our analysis by a classical utility optimization problem. Agents maximize  $u_i(E_i[f_i(m_i)], E_i[f_j(m_i)])$  with respect to their contribution  $m_i$ :

$$m_i \in \arg \max_{m_i} u_i(E_i[f_i(m_i)], E_i[f_j(m_i)]) \quad (13)$$

The first order condition of problem (13) reads

$$\frac{\partial u_i(E_i[f_i(m_i)], E_i[f_j(m_i)])}{\partial m_i} = 0, \quad (14)$$

with

$$\begin{aligned} & \frac{\partial u_i(E_i[f_i(m_i)], E_i[f_j(m_j)])}{\partial m_i} \\ &= (n-1) \cdot \left( \frac{\partial f_i(m_i, P_i(m_j))}{\partial m_i} - \frac{\partial f_j(m_i, P_i(m_j))}{\partial m_i} \right) \\ &= (-1 - k \cdot (1+r - n \cdot r)) \int_{m_i}^{\infty} P_i(m_j) dm_j \\ &+ k \cdot (1 - n + r) \cdot \int_0^{m_i} P_i(m_j) dm_j \cdot (n-1). \end{aligned} \tag{15}$$

The second-order condition for a local maximum of (13) holds for any reasonable assignment of the problem parameters, i.e.

$$\frac{\partial^2 u_i(E_i[f_i(m_i)], E_i[f_j(m_j)])}{\partial m_i^2} < 0, \forall k > 0, n \geq 2, 0 \leq g \leq n, r > 1, 0 < m_i < \infty.$$

The cumulative distribution function of the contributions  $m_j$  of the other agents, as anticipated by agent  $i$ , is defined by  $CDF_i(m_i) \equiv \int_0^{m_i} P_i(m_j) dm_j$ . The term  $\int_{m_i}^{\infty} P_i(m_j) dm_j$  in equation (15) corresponds to the survival function of the subjective expected distribution of contributions in the population:

$$a_i(m_i) := 1 - CDF_i(m_i) = P_i(\{m_j > m_i\}) = \int_{m_i}^{\infty} P_i(m_j) dm_j \tag{16}$$

Substituting  $a_i(m_i)$  as defined in equation (16) into equation (15) yields:

$$-1 + \frac{g}{n} + k \cdot (n-1) \cdot (a_i(m_i) \cdot r + a_i(m_i) - 1) = 0 \tag{17}$$

Equation (17) describes a functional relation between the predetermined parameters of the public goods game, i.e. the group size  $n$ , the project return factor  $g$  and the punishment efficiency  $r$ , as well as the variable traits of agent  $i$ , i.e. the propensity  $k$  to punish and her subjective expectation (first-order belief) about the fraction  $a_i(m_i)$  of her group fellows who contribute more than her own contribution  $m_i$ .

As we are interested in the agents' evolutionary optimal punishment behavior, we solve equation (17) for  $k$  and obtain:

$$k^* = \frac{1}{1 - n + r + a_i(m_i) \cdot (n-2) \cdot (1+r)} \tag{18}$$

$k^*$  depends on  $m_i$  via agent  $i$ 's subjective (first-order) belief embodied in  $a_i(m_i) \in [0,1]$  that the other agents will contribute more than herself. Thus,  $k^*$  can be interpreted as the value that makes agent  $i$  better off not to deviate negatively or positively from her willingness to contribute  $m_i$  MUs to the public good, given she believes a number of  $N = n \cdot a_i(m_i)$  of other group fellows contribute more than her own contribution  $m_i$ .

In the following subsection, we consider evolutionary dynamics in our model.

### 5 Evolutionary dynamics of the level of cooperation $m_i$ as function of the propensity to punish $k$

The evolutionary dynamics of agents, who face a social dilemma situation in the form of a public goods game with punishment, can be described in terms of the agents' P&L as a function of the deviation in the contribution level  $m_i(t)$  and of the population's common propensity to punish  $k$ . If agent  $i$  starts to deviate from her current level of cooperation  $m(t)$  by a value of  $\Delta m = m(t+1) - m(t)$ , the absolute change of the P&L for the agent as a function of  $\Delta m$  and  $k$  is defined as follows:

$$\Delta P\&L_i(\Delta m, k) = \begin{cases} g \cdot \frac{\Delta m}{n} - \Delta m + (n-1) \cdot k \cdot \Delta m \cdot r & , \Delta m \leq 0 \\ g \cdot \frac{\Delta m}{n} - \Delta m - (n-1) \cdot k \cdot \Delta m & , \Delta m > 0 \end{cases} \tag{19}$$

These expressions assume that the deviation affects a single agent, who deviates from the norm. Hence, the agent is punished by (resp. punishes) all other  $n-1$  agents for  $\Delta m < 0$  (resp.  $\Delta m > 0$ ).

The deviation of agent  $i$  by  $\Delta m$  affects not only her own P&L, but also the P&L of the remaining agents  $j = 1 \dots n, j \neq i$ . The absolute change of the P&L of the remaining population as a function of  $\Delta m$  and  $k$  reads

$$\Delta P\&L_j(\Delta m, k) = \begin{cases} g \cdot \frac{\Delta m}{n} + k \cdot \Delta m \cdot r & , \Delta m \leq 0 \\ g \cdot \frac{\Delta m}{n} - k \cdot \Delta m \cdot r & , \Delta m > 0. \end{cases} \tag{20}$$

Putting equations (19) and (20) together with

$$\tilde{\Delta P\&L}_i(\Delta m, k) := \Delta P\&L_i(\Delta m) - \Delta P\&L_j(\Delta m) \tag{21}$$

yields the relative change of the P&L of agent  $i$  with respect to the remaining population:

$$\tilde{\Delta P\&L}_i(\Delta m, k) = \begin{cases} -\Delta m + (n-1) \cdot k \cdot \Delta m \cdot r - k \cdot \Delta m & , \Delta m \leq 0 \\ -\Delta m - (n-1) \cdot k \cdot \Delta m + k \cdot \Delta m \cdot r & , \Delta m > 0 \end{cases} \tag{22}$$

The form of equation (22) is equivalent to the relative measure of success introduced in equations (3) and (7) with  $s := [m + \Delta m, k]$ . As introduced above, the realized P&L from the public goods game with punishment can be interpreted as the fitness of an agent in an evolutionary environment. The fitness, in turn, is associated with the rate of fertility, i.e. the fitter an agent becomes, the more genetically related offsprings she produces. In this way, traits of agents with a higher realized P&L value tend to spread and to end up dominating the population over time. It thus holds that the traits  $[m, k]$  in the population move with time towards values  $[\hat{m}, \hat{k}]$  of a subpopulation that on average achieves a higher mean P&L than the average mean P&L of the entire population.

The corresponding replicator dynamics are

$$\begin{aligned} \frac{\partial x(\Delta m)}{\partial t} &= \int_0^{\infty} \tilde{\Delta P\&L}(\Delta m, k) \cdot x(\Delta m) dk \\ \frac{\partial x(k)}{\partial t} &= \int_{-\infty}^{\infty} \tilde{\Delta P\&L}(\Delta m, k) \cdot x(k) d\Delta m. \end{aligned} \tag{23}$$

with  $x(\Delta m)$  and  $x(k)$  being respectively the proportion of agents deviating by  $\Delta m$  and with a propensity to punish  $k$ . Extending the previous reasoning, expressions (23) now include the occurrence of deviations from the optimal (for the community) propensity to punish, which may result for instance from random mutations. The dynamics for the expected group average,  $\bar{m}$  and  $\bar{k}$ , are accordingly defined by

$$\begin{aligned} \frac{\partial E[\bar{m}]}{\partial t} &= \int_{-\infty}^{\infty} \int_0^{\infty} \Delta m \cdot \tilde{\Delta P \& L}(\Delta m, k) \cdot x(\Delta m) \cdot x(k) dk d\Delta m \\ \frac{\partial E[\bar{k}]}{\partial t} &= \int_0^{\infty} \int_{-\infty}^{\infty} k \cdot \tilde{\Delta P \& L}(\Delta m, k) \cdot x(\Delta m) \cdot x(k) d\Delta m dk. \end{aligned} \tag{24}$$

The sensitivity of  $\tilde{\Delta P \& L}_i(\Delta m, k)$  given by (22) with respect to the relative change of  $\Delta m$  is defined by the partial derivative

$$\frac{\partial \tilde{\Delta P \& L}_i(\Delta m, k)}{\partial \Delta m} = \begin{cases} -1 - k + k \cdot (n-1) \cdot r & , \Delta m \leq 0 \\ -1 - k \cdot (n-1) + k \cdot r & , \Delta m > 0 \end{cases} \tag{25}$$

With the conditions that  $n \geq 2$  and  $r > 1$ , i.e. a game has always two or more players and punishment is less costly to the punisher than to the punished agent, it holds that, for  $\Delta m(t) > 0$ ,  $\frac{\partial \tilde{\Delta P \& L}_i}{\partial \Delta m}$  is always negative and, for  $\Delta m < 0$ ,  $\frac{\partial \tilde{\Delta P \& L}_i}{\partial \Delta m}$  is always positive:

- a)  $\frac{\partial \tilde{\Delta P \& L}_i(\Delta m, k)}{\partial \Delta m} \leq 0, \forall k \leq \frac{1}{n \cdot r - r - 1}, \Delta m < 0$
- b)  $\frac{\partial \tilde{\Delta P \& L}_i(\Delta m, k)}{\partial \Delta m} > 0, \forall k > \frac{1}{n \cdot r - r - 1}, \Delta m < 0.$

This reveals the existence of two distinct evolutionary regimes that are separated by the bifurcation point at

$$k^+ = \frac{1}{n \cdot r - r - 1}. \tag{27}$$

- **Defection:** For  $k \leq \frac{1}{n \cdot r - r - 1}$  and  $\text{Var}(m_j) > 0, j = 1, \dots, n$ , the linear P&L structure of the public goods game with punishment together with the replicator dynamics are responsible for  $\Delta m$  to become more negative over time. It intuitively follows that defection pays out, such that

$$m^a : = \lim_{t \rightarrow \infty} m(t) \approx \frac{c_{\text{fix}}}{g-1} \tag{28}$$

results as the evolutionary stable behavior. Remember that each agent has a minimum cost of living defined by  $c_{\text{fix}}$ . In order to meet this survival condition, the average minimum contribution of the population is constrained to values of  $m > \frac{c_{\text{fix}}}{g-1}$ .

- **Coordination:** For  $k > \frac{1}{n \cdot r - r - 1}$ , a heterogeneous population with  $\text{Variance}(m_j) > 0$  for  $j = 1, \dots, n$  follows a dynamic that

does not converge to a predetermined unique evolutionary attraction point but rather converges to an evolutionary stable set of strategies. As punishment is efficient in this regime, with  $\frac{\partial \tilde{\Delta P \& L}_i(\Delta m, k)}{\partial \Delta m} > 0$  for values of  $\Delta m < 0$ , the social dilemma problem transforms into a coordination problem [1]. If punishment is efficient, the utility maximizing strategy is to contribute according to the expected contribution of the remaining group fellows, i.e. to contribute according to the first-order belief. Following Black's theorem, the best estimate for this strategy is the median value  $\bar{m}_i$  of the subjective probability measure  $P_i$  that is believed to characterize the contributions of the group fellows [12,35–37]. The median value  $\bar{m}_i$  of the subjective probability distribution  $P_i$  is defined by

$$\int_{\bar{m}_i}^{\infty} P_i(m_j) dm_j = \frac{1}{2} \tag{29}$$

Consequently

$$m^b : = \lim_{t \rightarrow \infty} m(t) = \bar{m} \tag{30}$$

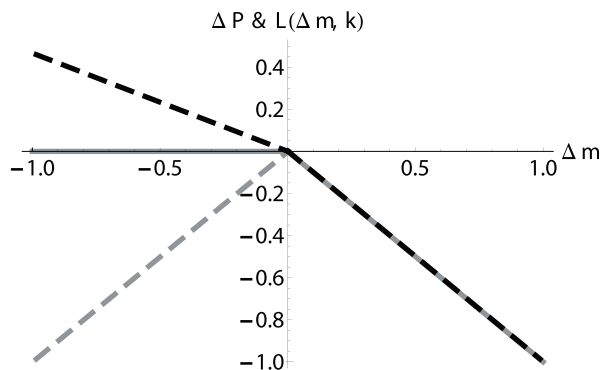
results as an evolutionary stable behavior in the population.

Figure 2 depicts the structure of equation (22) with a punishment efficiency factor  $r=3$  and a group size  $n=4$  for  $k = \frac{1}{15}$  (black, dashed),  $k=0.125$  (grey) and  $k=0.25$  (grey, dashed).

The next subsection analyzes the identified evolutionary stable strategies (ESSs) for a population of agents that is either purely self-regarding and acting selfishly or a population of agents that incorporates other-regarding preferences in their decision process.

### 6 Coevolutionary dynamics of the propensity to punish $k$ and the level of cooperation $m_i$ in the presence of self and other-regarding preferences

First, consider a population of purely self-regarding and selfish acting agents, i.e. agents who try to maximize their utility without e.g. taking into account specific preferences with respect to the P&L and the contributions of the remaining agents in the group. The preferences of self-regarding and selfish agents are simply



**Figure 2. Sensitivity of  $\tilde{\Delta P \& L}(\Delta m, k)$  as a function of a relative change  $\Delta m$  of the contributions for a group size of  $n=4$ , a punishment efficiency  $r=3$  and a propensity to punish of  $k = \frac{1}{15}$  (black, dashed),  $k=0.125$  (grey) and  $k=0.25$  (grey dashed).**  
doi:10.1371/journal.pone.0077041.g002



characterized by the dislike of situations in which their P&L in current period  $t$  is less than their P&L in the previous period  $t-1$ . This implies that all agents in the population are satisfied if and only if the following expression is fulfilled

$$\begin{aligned}
 f_i(m_1(t), \dots, m_n(t)) &\geq f_i(m_1(t-1), \dots, m_n(t-1)) \\
 \text{and} & \\
 f_j(m_1(t), \dots, m_n(t)) &\geq f_j(m_1(t-1), \dots, m_n(t-1))
 \end{aligned}
 \tag{31}$$

for all  $i=1, \dots, n$  with  $j=1, \dots, i-1, i+1, \dots, n$ . The functions  $f_i(\dots)$  and  $f_j(\dots)$  are defined in equation (5) and (6), respectively. Reducing the expression in (31) over the domain of reasonable values for the variables  $m_i \geq 0, m_j > 0, k \geq 0, 0 \leq a_i(m_i) \leq 1, n \geq 2, 0 < g < n$  and  $r > 1$  and solving it to the propensity to punish  $k$  gives the following condition:

$$k_{\text{self}} \geq \frac{n-g}{(n-1) \cdot n \cdot r}
 \tag{32}$$

The population of agents initially consists of uncooperative, non-punishers, i.e.  $k(t=0) \simeq 0$  and  $m_i(t=0) \simeq 0$  for  $i=1, \dots, n$ . For all reasonable values of  $n \geq 2, g \geq 0$  and  $r \geq 1$ , it holds that the desired propensity to punish of self-regarding agents is less than the bifurcation threshold  $k^+$ , defined in equation (27):

$$\frac{n-g}{(n-1) \cdot n \cdot r} \leq k_{\text{self}} < k^+
 \tag{33}$$

With the population being initialized at  $k(0) \simeq 0$  and the condition in equation (33), it follows that agents have an incentive to defect as can be inferred from equation (26a). In other words, agents have an incentive to contribute less than the amount contributed by the other group fellows. In general, agents have no ex-ante information about the others' contributions at the time they take the decision to contribute  $m_i$  MUs. However, agents have beliefs about the others' contribution that is embodied in the subjective probability distribution  $P_i$ . This allows them to form their expectations about the group average contribution as defined in equation (9). In terms of equation (16), "defecting" translates into a probability of one that all  $m_j$  values are larger than the own contribution  $m_i$ , i.e.  $a_i(m_j) = 1$ . With  $a_i(m_j) = 1$ , it follows that the optimal propensity to punish defined in equation (18) becomes

$$\begin{aligned}
 k_{\text{self}}^* &= \frac{1}{1-n+r+(n-2) \cdot (1+r)} \\
 &= \frac{1}{(n-1) \cdot r - 1}
 \end{aligned}
 \tag{34}$$

which is exactly equivalent to the evolutionary threshold value of  $k^+$  defined in equation (27). Plugging  $k_{\text{self}}^*$  into equation (22) yields

$$\tilde{\Delta P\&L}_i(\Delta m, k_{\text{self}}^*) = \begin{cases} 0 & , \Delta m \leq 0 \\ -\frac{\Delta m \cdot (n-2) \cdot (r+1)}{r \cdot (n+1) - 1} < 0 & , \Delta m > 0 \end{cases}
 \tag{35}$$

Together with the replicator dynamics defined in equation (24), it follows that the population converges towards the evolutionary stable attraction point for  $m^a$  that is defined in equation (28). For the evolutionary dynamics of  $k_{\text{self}}^*$ , it holds that

$$k_{\text{self}} \leq k_{\text{self}}^* \leq k^+,
 \tag{36}$$

i.e. the value range of the propensity to punish is restricted to the interval  $k(t) \in [k_{\text{self}}, k^+]$ . Thus, selfish and purely self-regarding agents are inevitably caught in the *defection* regime, as  $k_{\text{self}}$  does not allow to overcome the bifurcation hurdle at  $k^+$ . Consequently, the population converges towards the ESS that is defined by  $s_{\text{self}} = [m^a, k_{\text{self}}^*]$ .

Consider now a population of agents who display other-regarding behavior in the form of disadvantageous inequity aversion. In general, inequity aversion preferences relate the personal utility gained from a public good to the personal contributed effort. If an imbalance exists between the own contributed effort and the personally received payoff compared to the performed effort and the received payoff of other agents in the group, the outcome of the game is perceived as being inequitable or "unfair". Disadvantageous inequity aversion implies that subjects only dislike situations in which the inequity is to their disadvantage. The payoff of an agent  $i$ , who plays a public goods game with punishment, is defined by equation (5) and the personal effort is equivalent to the contributed amount of MU  $m_i$ . An agent with an aversion against disadvantageous inequitable outcomes thus does not like situations in which

- she contributes equally or more than her group fellows ( $m_i \geq m_j$ ) and receives a payoff that is smaller than the average utility received by the remaining group members ( $f_i < f_j$ ) or
- she contributes more to the public good ( $m_i > m_j$ ) and, at the same time, receives a payoff that is smaller or equal to the remaining group's utility ( $f_i \leq f_j$ ).

By implication, the population of agents is satisfied if, and only if, at least one of the following three conditions is fulfilled for all  $i=1, \dots, n$  with  $j=1, \dots, i-1, i+1, \dots, n$ :

- a)  $f_i(m_1, \dots, m_n) > f_j(m_1, \dots, m_n) \wedge m_i > m_j$ ,
- b)  $f_i(m_1, \dots, m_n) \geq f_j(m_1, \dots, m_n) \wedge m_i = m_j$ ,
- c)  $f_i(m_1, \dots, m_n) < f_j(m_1, \dots, m_n) \wedge m_i < m_j$ .

Expressing the above conditions (37) over the domain of eligible values for the variables  $m_i \geq 0, m_j \geq 0, k \geq 0, 0 \leq a_i(m_i) \leq 1, n \geq 2, 0 < g < n$  and  $r > 1$  and solving them in terms of the propensity to punish  $k$  yields the following condition:

$$k_{\text{ieq}} > \frac{1}{a_i(m_i) \cdot (r+1) \cdot (n-2) + r + 1 - n}
 \tag{38}$$

For all values  $0 \leq a_i(m_i) \leq 1$ , it holds that  $k_{\text{ieq}}$  in condition (38) becomes larger than the evolutionary threshold  $k^+$  defined in equation (27) such that

$$k^+ < \frac{1}{a_i(m_i) \cdot (r+1) \cdot (n-2) + r + 1 - n} < k_{\text{ieq}}, \forall 0 \leq a_i(m_i) \leq 1.
 \tag{39}$$

Thus, agents are forced to switch from the *defection* regime into the *coordination* regime in order to satisfy their preferences. As described before, the best response strategy in the coordination regime regarding the level of cooperation  $m_i$  is to contribute according to the median value of the subjective probability distribution  $P_i$ . By the definition in equation (16), it follows that the median value  $\bar{m}_i$

of  $P_i$  is equivalent to a value of  $a_i(\bar{m})=0.5$ . Plugging  $a_i(m_i)=0.5$  into equation (18) yields the following estimate for the optimal propensity to punish among disadvantageous inequity averse agents

$$k_{icq}^* = \frac{2}{n \cdot (r-1)}. \tag{40}$$

From that fact that  $k_{icq}^* > k^+$ , it follows that an evolutionary stable attraction point  $m^b$  emerges, which has been defined in equations (29) and (30), respectively. Substituting equation (40) into equation (22) results in a  $\Delta P\&L$  profile that is determined by symmetrically downward sloping functions  $\tilde{\Delta P\&L}_i(\Delta m, k^b)$  centered relative to the maximum at  $\Delta m = 0$  with

$$\tilde{\Delta P\&L}_i(\Delta m, k^b) = \begin{cases} \frac{\Delta m \cdot (n+2) \cdot (r+1)}{r \cdot (n-1) - 1} < 0 & , \Delta m \leq 0 \\ -\frac{\Delta m \cdot (n+2) \cdot (r+1)}{r \cdot (n-1) - 1} < 0 & , \Delta m > 0 \end{cases}. \tag{41}$$

The population of agents is thus able to maintain a stable level of cooperation at the median value  $\bar{m}$  that is determined by the initial set of subjective expectations about the other contributions embodied in the distributions  $P_i$ .

In conclusion, a population of disadvantageous inequity averse agents converges to the ESS that is determined by  $s_{icq} = [m^b, k_{icq}^*]$ .

Our first main result can be summarized as follows:

**Result 1:** *In the presence of standard Darwinian evolutionary dynamics, agents' traits (strategies) converge to evolutionary stable strategies, which results in a public goods game with punishment to be either characterized by defection (for weak punishment) or by coordination (for sufficient punishment). Purely self-regarding agents are inevitably caught in the defection regime while disadvantageous inequity averse agents are able to resolve the social dilemma by transforming it into a coordination problem.*

In the following section, we turn to the empirical validation of our model.

### Empirical Test of the Theory

In this section, we compare the predictions derived from our model with the empirical data obtained in three independently conducted lab experiments and validate our results against the empirical observations.

#### 1 Description of the empirical data set

We analyze data from three public goods game experiments with punishment [19–21], which were carried out independently. In each experiment, groups of  $n=4$  subjects played a two-stage public goods game: at the beginning of stage one, the contribution step, individuals were endowed with 20 monetary units (MUs). Subjects could decide on the amount  $m_i$  of MUs to contribute to the public good. The sum of all contributions was compounded by a factor of  $g=1.6$  and subsequently redistributed in equal shares to all group members. Note that this results in a per capita gain of  $0.4 < 1$  per contributed MU, which induced a distinct social dilemma component. In the second stage, the punishment step, subjects were informed about the contributions of their group mates. Subsequently, they could spend an additional fraction of their endowment to punish other group fellows. Each MU spent by the punisher caused a harm of approximately  $r=3$  MUs to the punished subject. The punishment step in the first experiment was slightly different from the other two experiments: the punishment efficiency factor was determined based on the first stage payoff of

the punished individual. However, it can be considered to be approximately equal to the factor 3 as in the remaining two experiments.

These two stages were played repetitively either in a stranger or a partner treatment. In the former, group members were reshuffled after each iteration to preserve the characteristics of one-shot interactions, i.e., to control for direct reciprocal effects. In the partner treatment, subjects played continuously with the same group members across all periods. The first experiment was composed of both a stranger and a partner treatment. Each of them were played for 10 periods. The second and third experiments included only a stranger treatment and were played for 6 and 10 iterations, respectively. In addition, the third experiment differed in the way information about the received punishment was revealed to the punished subjects. In the first one, the so-called observed treatment, subjects were informed immediately after the punishment stage about the costs of the received punishment, as in experiments one and two. In contrast, in the second treatment, the unobserved treatment, subjects were informed about the costs they had to bear for being punished only after the last period had been played. However, the results of both treatments were found not to be significantly different as the fear of punishment seems to be as effective as the punishment itself [21]. To obtain a sufficiently large sample size, we pool the observations from all treatments of the three experiments introduced above. The subject pool size amounts to a total of 440 subjects.

#### 2 Recovering the propensity to punish from the empirical data

The empirical propensity to punish can be calculated by taking the observed deviations  $(m_i - m_j) > 0$  between subject  $i$  and  $j$  and the observed punishment from subject  $i$  to  $j$ ,  $p_{i \rightarrow j}$ . In this way, each pairwise interaction between two subjects provides a realization for the propensity to punish according to the formula

$$k_{ij} = \frac{p_{i \rightarrow j}}{m_i - m_j}. \tag{42}$$

With the set of all pairwise interactions, we construct the empirical distribution of the propensity to punish, by sampling all realized  $p_{i \rightarrow j}$  with their corresponding  $m_i$  and  $m_j$ .

As shown in the first section and also demonstrated in [38], the agents' propensity to punish can be interpreted as a norm-enforcing behavior that has co-evolved over tens and hundreds of thousands of years by gene-culture co-evolution along with the emergence of an aversion to disadvantageous inequitable outcome. The perception of fairness and the reaction to unfair behavior seems to be deeply rooted in our cultural and genetic heritage [28,39], as experiments and field studies across different locations and cultural groups suggest [15,31]. We thus consider the propensity to punish  $k$  to be a constant on the evolutionary negligible short time-scale of the experiments. This can be substantiated by comparing the results of a two-sample Kolmogorov-Smirnov test between an empirical data set containing only data from the first period and the corresponding full-sample data set. The null hypothesis that the distributions of the two data sets of  $k_{ij}$  result from the same generating mechanism cannot be rejected ( $p$ -value equal to 0.31). In all three experiments, the observed contributions  $m_i > 0$  are approximately stable over time, as they do not converge towards full defection. Additionally, the standard deviation of the contributions is on average decreasing



over time. Both of these measures indicate that the subjects in the experiments are in the “*coordination*” regime.

### 3 Validation of the model prediction for $k$

We validate the model by asking whether the ESS value  $k_{ieq}^*$  of the propensity to punish in the *coordination* regime given by equation (40) matches the empirically observed data. The group size  $n$  and punishment efficiency  $r$  are known parameters in the experiments. The three public goods game experiments with punishment [19–21] have been performed with  $n=4$  players and a punishment efficiency factor of  $r=3$ , respectively. Plugging both values into equation (40) yields

$$k_{ieq}^* = \frac{1}{4}. \tag{43}$$

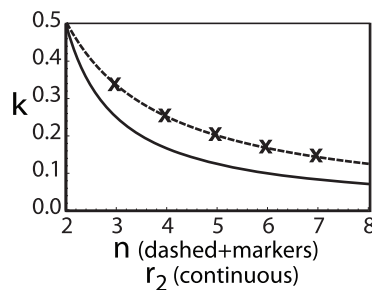
As the value given by (43) is based on the assumption that subjects contribute according to the median value of their subjective probability distribution about the contributions of their group fellows,  $k_{ieq}^*$  corresponds consequently to the median of the distribution of the values  $\{k_{i \rightarrow j}\}$  of the propensity to punish.

Remarkably, we find an exact match with the median value  $\tilde{k}_{emp}$  estimated from the empirical distribution of the  $\{k_{i,j}\}$  values, i.e.  $k_{ieq}^* = \tilde{k}_{emp} = 0.25$ . The standard error of the median of the empirical data is  $\hat{\sigma}_{med}^k = 0.0013$ . This corresponds to a one-standard error range given by  $\tilde{k}_{emp} \pm \hat{\sigma}_{med}^k = [0.2487, 0.2513]$ . The corresponding 95% confidence intervals for the sample median values are  $CI_{0.95}^* = [0.2423, 0.2655]$ ,  $CI_{0.95}^+ = [0.2486, 0.3336]$ ,  $CI_{0.95}^- = [0.1568, 0.2611]$  and  $C_{0.95}^{all} = [0.25, 0.25]$ . To estimate the confidence intervals we used a bootstrap t-method presented in [40].

The superscript on the *CI* indicates the individual data sets:

- $\star =$  [19]
- $+$  = [20]
- $-$  = [21]
- $all$  = pooled data set of all three experiments.

This remarkable agreement between theory and empirical data suggests that subjects act according to the optimization problem defined in (13) and that their punishment behavior is dominated by disadvantageous inequity aversion preferences defined in equation (37). Again, we argue that in this specific setup the focal action to punish negative deviators by spending roughly a fourth of the negative deviation has emerged as the result of the human’s



**Figure 3. Propensity to punish as a function of the punishment efficiency  $r$  (continuous line) for a fixed group size  $n=4$  and as a function of the group size  $n$  (dashed line with cross markers) for a fixed  $r=3$ .**

doi:10.1371/journal.pone.0077041.g003

psychological predisposition to render effective the culturally and genetically internalized norms [11,38]. In this case, these norms are described by disadvantageous inequity aversion. We can now state our second main result:

**Result 2:** *The level of altruistic punishment that subjects exhibit in public goods game experiments can be explained by a simple aversion to disadvantageous inequitable outcomes together with the individual maximization of the expected utility defined in equation (13) and the presence of evolutionary dynamics.*

The dependence of the optimal propensity to punish  $k_{ieq}^*$  defined in equation (40) on the group size  $n$  and the punishment efficiency factor  $r$  is plotted in figure 3. This predicts the potential propensity to punish that should be observed in experiments with differing configurations. In particular, the larger the punishment efficiency  $r$  and the group size  $n$ , the smaller becomes the optimal propensity to punish. To validate these predictions additional experiments with different groups sizes and punishment efficiency factors have to be performed in future research.

The following section analyzes the co-evolutionary dynamics of agents with disadvantageous inequity aversion compared to agents with purely self-regarding and selfish behavior in a heterogeneous population.

### Evolutionary Dominance of Other-Regarding Preferences

The results and findings presented in the previous two sections inevitably raise the question about the evolutionary stability and dominance of other-regarding compared to self-regarding preferences. Are agents with other-regarding behavior able to invade a population of initially selfish and self-regarding agents? Can the required conditions for the emergence of altruistic punishment spread in a population of agents that is facing a competitive resource limited environment as described by our model? Is disadvantageous inequity aversion the predominant strategy in a population of agents who face a social dilemma situation that provides the opportunity to punish? This section addresses these questions by providing an analysis of the co-evolutionary dynamics that are at play in a heterogeneous population consisting of a mixture of disadvantageous inequity averse agents and purely self-regarding and selfish-acting agents.

A system that is subject to evolutionary forces is characterized and determined by selection, cross-over and mutation processes. Consequently, the birth and death of agents induce multifaceted and complex co-evolutionary dynamics that are contingent on the states and path dependencies of the individual actors in the system. In view of this complexity, this section presents a simplified but conclusive analytical representation of the system’s dynamics and properties. This is achieved by reducing the assumed heterogeneity in the system and by considering only two groups and types of agents, respectively. An extensive numerical analysis of a population of agents playing a public goods game with punishment that takes into account the full heterogeneity and the full set of evolutionary dynamics and path dependencies is presented in the companion article [38].

#### 1 Conditions for evolutionary dominance

Let us write the evolutionary success of a homogeneous group **A** of agents with size  $d$  playing strategy  $s_1 = [m_1, k_1]$  that competes with a homogeneous group **B** of size  $n-d$  with agents playing strategy  $s_2 = [m_2, k_2]$ . Using equation (3) and the P&L structure of the public goods game with punishment defined in the equations (5,6), we obtain

$$\begin{aligned}
 \Phi(d, n, s_1, s_2) &= \sum_d f_1(m_1, \dots, m_n) - \sum_{n-d} f_2(m_1, \dots, m_n) \\
 &= - \sum_d m_1 + \sum_{n-d} m_2 - \sum_d \sum_{n-d} k_1 \cdot \max(m_1 - m_2, 0) \\
 &\quad + \sum_d \sum_{n-d} k_1 \cdot \max(m_1 - m_2, 0) \cdot r \\
 &\quad - \sum_{n-d} \sum_d k_2 \cdot \max(m_2 - m_1, 0) \cdot r \\
 &\quad + \sum_{n-d} \sum_d k_2 \cdot \max(m_2 - m_1, 0) \\
 &\quad + g \cdot \frac{d - (n-d)}{n} \cdot [dm_1 + (n-d)m_2].
 \end{aligned} \tag{44}$$

The measure  $\Phi^+$  defines a relation between the relative difference of the P&L of group **A** versus that of group **B**. It thus reflects the evolutionary success or failure of the two competing groups over time. An expected deviation of group **A** by a value of  $\Delta \hat{m}$  affects  $\Phi^+$  to become either positive or negative. Depending on the sign of  $\Phi^+$ , either the strategies of group **A** start to dominate the population ( $\Phi^+ > 0$ ) or alternatively, if  $\Phi^+ < 0$ , the strategies of group **B** spread and dominate in the population.

## 2 Evolutionary dominance of disadvantageous inequity averse agents

Consider a population of size  $n$  that initially consists only of purely self-regarding and selfish acting agents. This homogeneous population is assumed to be in an evolutionary equilibrium state. As identified in the previous sections, self-regarding agents play the ESS  $s_{\text{self}} = [m^a, k_{\text{self}}^*]$  with

$$k_{\text{self}}^* = \frac{n-g}{(n-1) \cdot n \cdot r}$$

and

$$m^a \approx \frac{c_{\text{fix}}}{g-1}$$

as given by the equations (32) and (28). Replacing one agent in this population by a disadvantageous inequity averse agent leads to a heterogeneous population that consists of two homogeneous subgroups. In the following, we analyze the co-evolutionary dynamics of this heterogeneous population of agents that is composed of a group **A** with size  $n-1$  of purely self-regarding agents and a group **B** with a single disadvantageous inequity averse agent corresponding to size  $d=1$ .

In contrast to the self-regarding agents, disadvantageous inequity averse agents play the ESS given by  $s_{\text{ieq}} = [m^b, k_{\text{ieq}}^*]$  with expression (40) for

$$k_{\text{ieq}}^* = \frac{2}{n \cdot (r-1)}$$

and

$$m^b = \bar{m}$$

as defined in equations (40) and (30).

We need now to specify the punishment propensities  $k_1$  and  $k_2$  that enter into expression (45). For this, we take the view that agents are boundedly rational and choose their punishment propensity using a model of others based on homophily or, more precisely, a theory of minds that attribute to others one's own inclinations. This amounts to choosing  $k_1 = k_{\text{ieq}}^*$  given by (40),  $k_2 = k_{\text{self}}^*$  given by (34). Reporting these expressions together with  $d=1$  into equation (45), we obtain

$$\begin{aligned}
 \Phi^*(1, n, p_1, k_{\text{ieq}}^*, k_{\text{self}}^*) &= \frac{(p_1 - 1) \cdot (2 - n) \cdot \Delta m p_1 \cdot (n - g) \cdot \Delta m}{n \cdot r} + \\
 &\quad \frac{(g \cdot (2 - 3p_1 + n \cdot (2p_1 - 1)))}{n}.
 \end{aligned} \tag{46}$$

The logically consistent relation between the evolutionary success or failure, viewed either from the perspective of group **A** or from group **B**, reads:

$$\begin{aligned}
 \Phi_{\mathbf{B}}^* &:= \underbrace{\Phi^*(1, n, p_1, k_{\text{ieq}}^*, k_{\text{self}}^*)}_{\text{perspective of group B}} = \\
 &\quad - \underbrace{\Phi^*(1, n, 1 - p_1, k_{\text{self}}^*, k_{\text{ieq}}^*)}_{\text{perspective of group A}} =: -\Phi_{\mathbf{A}}^*
 \end{aligned} \tag{47}$$

If  $\Phi_{\mathbf{B}}^* > 0$ , group **B** that initially consists of a single disadvantageous inequity averse agent, outperforms group **A** that has  $n-1$  members of self-regarding agents. Consequently, the strategy  $s^b = [m^b, k_{\text{ieq}}^*]$  spreads in the population. In contrast, if  $\Phi_{\mathbf{A}}^* > 0$ , group **A** becomes predominant and strategy  $s^a = [m^a, k_{\text{self}}^*]$  spreads in the population. The resulting condition for the disadvantageous inequity aversion trait to become dominant is defined by

$$\Phi_{\mathbf{B}}^* > 0 \wedge \Phi_{\mathbf{A}}^* < 0. \tag{48}$$

Reducing condition (48) over the set of reasonable parameter values with  $\Delta m > 0$ ,  $n \geq 2$ ,  $r > 1$  and  $0 < g < n$  reveals that  $\Phi_{\mathbf{B}}^*$  becomes positive if the probability  $p_1$  for deviating negatively falls into the range

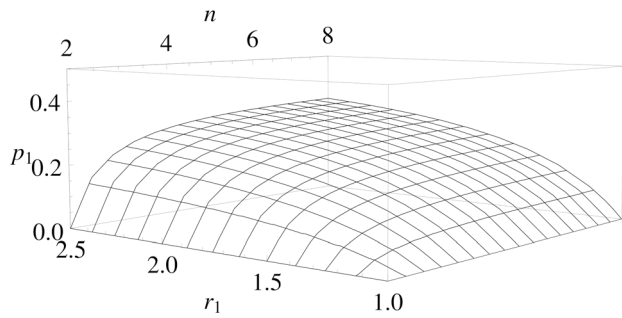
$$p_{\text{low}} < p_1 \leq 1 \tag{49}$$

with

$$p_{\text{low}} = \frac{(n-2) \cdot (g-1)}{2-3 \cdot g+n \cdot (2 \cdot g-1)}. \tag{50}$$

Figure 4 shows the surface defined by expression (49) for  $p_{\text{low}}$  as a function of  $n$  and  $r$  in the range  $2 < n < 8$  and  $1 < r < 2.5$ . The domain above the surface corresponds to  $p_1$  values for which a single disadvantageous inequity averse agent can invade a population of selfish agents by deviating from the contribution of the selfish agents.

A scenario with a population consisting of 4 agents with 3 agents being self-regarding and one agent being disadvantageous inequity averse, playing a public goods game with a per capita return of 0.4 MUs per invested MU, i.e.  $g=1.6$ , results in a  $1-p_{\text{low}}=82\%$  chance for the single disadvantageous inequity averse agent to outperform at each period.



**Figure 4. Minimum probability threshold  $p_1$  given by expression (49), above which a single disadvantageous inequity averse agent can invade a population of selfish agents by deviating from the contribution of the selfish agents with  $\Delta\hat{m} = p_1 \cdot (-\Delta m) + (1 - p_1) \cdot \Delta m$ .**  
doi:10.1371/journal.pone.0077041.g004

For all reasonable parameter values,  $n > 2$  and  $0 < g < n$ , the lower bound  $p_{low}$  is always smaller than  $\frac{1}{2}$ . This means that the probability for the disadvantageous inequity averse agent to invade the population of selfish agents over time is always larger than one-half. The range of  $p_1$  defined by equation (49) shows that, if the single disadvantageous inequity averse agent in group **B** deviates on average by a negative value, i.e.  $p_1 > \frac{1}{2}$ , from the contribution  $m_2$  of the selfish agents (group **A**), she always wins since  $\Phi_{\mathbf{B}}^* > 0$ .

Such a single agent can win even though she may be strongly out-numbered by the  $n-1$  selfish agents who tend to defect, because the minimum required consumption  $c_{fix}$  per period forces the population to contribute on average at least an amount of

$$\frac{d \cdot m_1 + (n-d) \cdot m_2}{n} \approx \frac{c_{fix}}{g-1}$$

MUs in order not to go extinct.

On the other hand, if the single disadvantageous inequity averse agent contributes on average more than the group of self-regarding and selfish agents, it must hold that

$$\frac{g+n-2}{2-3 \cdot g+n(2 \cdot g-1)} < \Delta\hat{m} \quad (51)$$

in order for that agent to have a larger P&L than the self-regarding agents of group **A**. Coming along with the condition  $\Phi_{\mathbf{B}}^* > 0$ , the disadvantageous inequity averse agent in group **B** can be thought of as being more fertile than the self-regarding agents of group **A**, which results in  $d(t+1)$  being larger than  $d(t)$  over time. In addition, with an increasing number  $d$  of agents in group **B** and, consequently, a decreasing number  $n-d$  of agents in group **A**, the lower limit for  $p_1$  declines until it becomes zero for  $d = \frac{n}{2}$ . This means that, as soon as half of the total population consists of disadvantageous inequity averse agents, the self-regarding and selfish agents are doomed, as the probability for group **B** to take over the entire population becomes 1 independent of their contribution decisions. These results are similar to those obtained by [41,42] on the “repeated prisoner’s dilemma” game and variants, for which larger minimum usage frequencies are needed to stabilize strategies with lower efficiency or payoffs.

In summary, for arbitrary initial conditions, we have established that disadvantageous inequity averse preferences and the corre-

sponding ESS  $s_{ieq}$  have significantly more than 50% chance of spreading in the population. At large times and for finite populations, in the presence of a larger than 50% probability to grow their relative population ( $1 - p_{low} > \frac{1}{2}$ ), the population of the disadvantageous inequity averse agents will with probability one reach half the total population, at which point they invade with certainty the whole population due to their self-reinforcing advantage explained above. This can be summarized by the following set of inequalities:

$$\begin{aligned} 1 - p_{low} > \frac{1}{2} &\Rightarrow \Pr[\Phi_{\mathbf{B}}^*(t) > 0] > \frac{1}{2} \\ &\Rightarrow \Pr[d(t+1) > d(t)] > \frac{1}{2} \Rightarrow \Pr[\Phi_{\mathbf{B}}^*(t+1) > \Phi_{\mathbf{B}}^*(t)] > \frac{1}{2} \quad (52) \\ &\Rightarrow \lim_{t \rightarrow \infty} \Pr[\Phi_{\mathbf{B}}^*(t) > 0] = 1 \Rightarrow \lim_{t \rightarrow \infty} d(t) = n. \end{aligned}$$

In conclusion, our third main result can be summarized as follows:

**Result 3:** *On long enough time scales, disadvantageous inequity averse preferences always invade and dominate pure self-regarding and selfish preferences in an evolutionary system.*

## Discussion

Prosocial behavior in humans has been studied in many laboratory experiments throughout the world. One key finding is the evidence for altruistic punishment behavior in humans, i.e. the punishment of non-cooperators and norm violators at own costs without direct or indirect material benefit [20,21,27,43–49]. To allow for this pro-social behavior that is often marked as “irrational”, researchers shifted from purely self-regarding assumptions to theories that incorporated other-regarding preferences [50]. In particular, analytical frameworks of fairness, reciprocity and cooperation have been formulated that combine individual utility maximization with inequality and inequity aversion [1–3,51–54]. These frameworks have a descriptive character and aim at reproducing the observed empirical behavior either quantitatively or using stylized facts. In this way, results from experimental economics have been rationalized and aligned with the predominant rational choice theory of pure self-interest.

Besides these equilibrium-based and time-independent utility theories, a second class of models emerged that focuses on the evolutionary origin of altruistic punishment and cooperation [55–62]. These models are often motivated from a biological perspective including arguments from evolutionary psychology, anthropology and sociology. Although the emergence of pro-social behavior in settings which are subject to material self-interest seems to contradict rational choice theory and the principle of the survival of the fittest, it is possible to show that cooperation and altruistic punishment can emerge and can be sustained in competitive, resource-limited environments [38,63–66]. Let us also mention studies of spatial versions of public good games with punishment [67,68], which have found that spatially evolving segregation between different cooperator and non-cooperator types may favor the emergence of cooperation. Contrary to the descriptive utility frameworks the evolutionary approaches aim at providing insight into the generating mechanisms of altruistic punishment and cooperation on an different level of analysis, and therefore are vague on what the exact nature of our social preferences should be. In the majority of cases, this is done by reproducing stylized facts rather than providing an external

validity of the empirically observed behavior as it is the case for most of the descriptive utility frameworks.

The lack of a close connection between the evolutionary literature on cooperation and altruistic punishment and the experimental literature has led to intense discussion and diverse interpretations on what the experimental results show and do not show [69–74]. This paper has tried to fill the gap between the evolutionary theoretical literature on cooperation and punishment and the empirical findings from experimental economics by combining both worlds: an expected utility framework that allows for standard evolutionary dynamics in the form of adaptation and selection. In particular, we showed that the interplay of natural selection and adaptation by selfish utility maximization inevitably results in the evolutionary dominance of other-regarding preferences in the form of disadvantageous inequity aversion when competing with pure self-regarding behavior. The term “disadvantageous” implies a relaxation from the concept of inequity aversion and fairness preferences: Subjects only dislike situations in which the inequity is to their disadvantage. Consequently, no a priori stipulated modeling assumptions about altruistic and self-discriminating behavior are embodied. The aversion against inequitable outcomes causes altruistic punishment behavior to emerge, even in social dilemma situations that are subject to material self-interest. We have argued that the bare individual survival needs of our ancestors induced an inherent predisposition to unfairness aversion that persists in our behavior up to this day.

This argument might sound farfetched given that human beings are probably the most successful species in eluding or manipulating natural selection by continuous enhancing, e.g., via improvements of health care and medical engineering. However, at the same time, our cultural evolution developed higher, more abstract levels of selection mechanisms that operate e.g. as monetary, bargaining and market competition, and led to hierarchical structures of power and of social standing. In other words, the natural selection that was previously affecting and operating on our hunter-gatherer ancestors has substantially been replaced in our modern societies by social institutions, most notably by the advent of money and the measures of economic power. Our primal instinct to unfairness aversion is still subliminally active and can be triggered by this high-order social and cultural selection mechanisms. In consequence, the corresponding reactions to unfair behavior can be observed today even though we are in most situations not directly affected in our biological viability.

## References

- Fehr E, Schmidt KM (1999) A Theory Of Fairness, Competition, And Cooperation. *The Quarterly Journal of Economics* 114 (3): 817–868.
- Bolton GE, Ockenfels A (2000) ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review* 90 (1): 166–193.
- Rabin M (1993) Incorporating Fairness into Game Theory and Economics. *The American Economic Review* 83 (5): 1281–1302.
- Robson AJ (1990) Efficiency in Evolutionary Games: Darwin, Nash, and the Secret Handshake. *Journal of Theoretical Biology* 144: 379–396.
- Robson AJ (1996) A Biological Basis for Expected and Non-expected Utility. *Journal of Economic Theory* 68: 397–424.
- Curry PA (2001) Decision Making under Uncertainty and the Evolution of Interdependent Preferences. *Journal of Economic Theory* 98: 357–369.
- Ely JC, Yilankaya O (2001) Nash equilibrium and the evolution of preferences. *Journal of Economic Theory* 97: 255–272.
- Samuelson L (2001) Analogies, Adaptation, and Anomalies. *Journal of Economic Theory* 97: 320–366.
- Samuelson L, Swinkels J (2006) Information, Evolution and Utility. *Theoretical Economics* 1: 119–142.
- Dekel E, Ely JC, Yilankaya O (2007) Evolution of Preferences. *Review of Economic Studies* 74: 685–704.
- Gintis H (2009) *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press.
- Bernheim BD (1994) A Theory of Conformity. *The Journal of Political Economy* 102 (5): 841–877.
- Messick D (1999) Alternative logics for decision making in social settings. *Journal of Economic Behavior & Organization* 39 (1): 11–28.
- Bardsley N, Sausgruber R (2005) Conformity and reciprocity in public good provision. *Journal of Economic Psychology* 26 (5): 664–681.
- Henrich J (2004) Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization* 53 (1): 3–35.
- Oechssler J, Riedel F (2001) Evolutionary dynamics on infinite strategy spaces. *Economic Theory* 17 (1): 141–162.
- Cressman R, Hofbauer J (2005) Measure dynamics on a one-dimensional continuous trait space: theoretical foundations for adaptive dynamics. *Theoretical Population Biology* 67 (1): 47–59.
- Hofbauer J, Oechssler J, Riedel F (2009) Brown-von Neuman-Nash dynamics: The continuous strategy case. *Games and Economic Behavior* 65 (2): 406–429.
- Fehr E, Gächter S (2000) Cooperation and Punishment in Public Goods Experiments. *The American Economic Review* 90 (4): 980–994.
- Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* 415 (6868): 137–140.
- Fudenberg D, Pathak PA (2010) Unobserved punishment supports cooperation. *Journal of Public Economics* 94:78–86.
- von Neumann J, Morgenstern O (2007) *Theory of Games and Economic Behavior* (Commemorative Edition) (Princeton Classic Editions), 60th Edition. Princeton University Press.
- Simon H, Egidi M, Viale R, Marris RL (2007) *Economics, Bounded Rationality and the Cognitive Revolution*. Edward Elgar Pub.

Previous works on economic theories about fairness, altruistic punishment and cooperation in voluntary contribution situations have systematically underestimated the importance of evolutionary dynamics and in particular the role of natural selection for the emergence of prosocial behavior and fairness preferences. We have combined an evolutionary approach together with an expected utility model to identify and explain the mechanisms that account for the emergence of fairness preferences and altruistic punishment. In particular, we designed an evolutionary utility model that allowed us to calculate an optimal strategy profile for the level of punishment in public goods games, depending on the fairness preferences of the agents in the population.

We considered two specific types of agents: (1) purely self-regarding agents and (2) agents who are disadvantageous inequity averse. We find that the evolutionary optimal strategy profile of disadvantageous inequity averse agents matches the behavior of subjects in the experiments and explains quantitatively the observed level of altruistic punishment without adjustable parameters. Our results imply that subjects show a strong predisposition for disadvantageous inequity aversion which, in turn, seems to be the driving force behind the observed altruistic punishment behavior. Finally, we showed that disadvantageous inequity aversion is an evolutionary dominant and stable strategy when compared to the pure self-regarding behavior, in a heterogeneous population of agents. Our theory offers new predictions that are testable by running future experiments with different numbers of subjects, modified payoff levels or a varied efficiency of the punishment.

In conclusion, we believe that path-dependent evolutionary processes, together with the self-organizational aspects of individual utility maximization, provide an important explanatory basis for the emergence of cooperation, altruism and prosocial behavior in general. Future research on social preferences should therefore include and focus on the time dimension and the evolutionary dependencies as well as the self-organizational capabilities of many social system.

## Author Contributions

Analyzed the data: MH DS. Wrote the paper: MH DS. Formulated the problem: MH DS. Developed the theory: MH DS. Analyzed the experimental data: MH DS.

24. Arthur WB (1994) Inductive Reasoning and Bounded Rationality. *The American Economic Review* 84 (2): 406–411.
25. Gigerenzer G, Selten R (2002) *Bounded Rationality: The Adaptive Toolbox*. The MIT Press.
26. Fehr E, Gächter S (2005) Human behaviour: Egalitarian motive and altruistic punishment (reply). *Nature* 433 (7021).
27. Egas M, Riedl A (2008) The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences* 275 (1637): 871–878.
28. Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, et al. (2001) In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review* 91 (2): 73–78.
29. Bowles S, Gintis H (2004) The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology* 65 (1): 17–28.
30. Gintis H (2003) The Hitchhiker's Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms. *Journal of Theoretical Biology* 220 (4): 407–418.
31. Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, et al. (2006) Costly Punishment Across Human Societies. *Science* 312 (5781): 1767–1770.
32. Henrich J, Ensminger J, McElreath R, Barr A, Barrett C, et al. (2010) Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. *Science* 327 (5972): 1480–1484.
33. Gould SJ, Lewontin RC (1979) The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proceedings of the Royal Society of London, Series B* 205: 581–598.
34. Savage IJ (1972) *The Foundations of Statistics*, 2nd Edition. Dover Publications.
35. Black D (1948) On the Rationale of Group Decision-making. *The Journal of Political Economy* 56 (1): 23–34.
36. Arrow KJ (1970) *Social Choice and Individual Values*, Second edition (Cowles Foundation Monographs Series), 2nd Edition. Yale University Press.
37. Selten R, Ostromann A (2000) *Imitation Equilibrium*. Tech. Rep. bgse16 2000, University of Bonn, Germany.
38. Hetzer M, Sornette D (2013) The co-evolution of fairness preferences and altruistic punishment. *PLoS ONE* 8 (8): e54308, pp.1–18, doi:10.1371/journal.pone.0054308.
39. Gintis H, Bowles S, Boyd R, Fehr E (2003) Explaining altruistic behavior in humans. *Evolution and Human Behavior* 24 (3): 153–172.
40. Efron B, Tibshirani RJ (1994) *An Introduction to the Bootstrap* (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), 1st Edition. Chapman and Hall/CRC.
41. Bendor J, Swistak P (1995) Types of evolutionary stability and the problem of cooperation. *Proceedings Of The National Academy Of Sciences Of The United States Of America* 92(8): 3596–3600.
42. Bendor J, Swistak P (1997) The Evolutionary Stability of Cooperation. *American Political Science Review* 91 (2): 290–307.
43. Bochet O, Page T, Putterman L (2006) Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization* 60 (1): 11–26.
44. Nikiforakis N, Normann HT (2008) A comparative statics analysis of punishment in public-good experiments. *Experimental Economics* 11 (4): 358–369.
45. Nikiforakis N (2010) Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior* 68: 689–702.
46. Anderson CM, Putterman L (2006) Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior* 54 (1): 1–24.
47. Brandts J, Fernanda Rivas M (2009) On punishment and well-being. *Journal of Economic Behavior & Organization* 72 (3): 823–834.
48. Gächter S, Renner E, Sefton M (2008) The Long-Run Benefits of Punishment. *Science* 322 (5907): 1510+.
49. Masclet D, Noussair C, Tucker S, Villeval MC (2003) Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. *The American Economic Review* 93 (1): 366–380.
50. Camerer CF (2003) *Behavioral Game Theory: Experiments in Strategic Interaction* (Roundtable Series in Behavioral Economics). Princeton University Press.
51. Cox JC, Friedman D, Sadiraj V (2008). Revealed Altruism. *Econometrica* 76 (1): 31–69.
52. Falk A, Fischbacher U (2006) A theory of reciprocity. *Games and Economic Behavior* 54 (2): 293–315.
53. Englmaier F, Wambach A (2010) Optimal incentive contracts under inequity aversion. *Games and Economic Behavior* 69 (2): 312–328.
54. Andreoni J, Miller J (2002) Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica* 70 (2): 737–753.
55. Axelrod R., Hamilton WD (1981) The Evolution of Cooperation. *Science* 211 (4489): 1390–1396.
56. Bowles S (1998) The Moral Economy of Communities Structured Populations and the Evolution of Pro-Social Norms. *Evolution and Human Behavior* 19 (1): 3–25.
57. Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. *Nature* 437 (7063): 1291–1298.
58. Lehmann L, Keller L (2006) The evolution of cooperation and altruism—a general framework and a classification of models. *Journal of evolutionary biology* 19 (5): 1365–1376.
59. Nowak MA (2006) Five rules for the evolution of cooperation. *Science* 314 (5805): 1560–1563.
60. Jensen K (2010) Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1553): 2635–2650.
61. Gächter S, Herrmann B, Thoeni C (2010) Culture and cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1553): 2651–2661.
62. Berger U (2011) Learning to cooperate via indirect reciprocity. *Games and Economic Behavior* 72: 30–37.
63. Imhof LA, Fudenberg D, Nowak MA (2005) Evolutionary cycles of cooperation and defection. *Proceedings of the National Academy of Sciences of the United States of America* 102 (31): 10797–10800.
64. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K (2007) Via freedom to coercion: The emergence of costly punishment. *Science* 316 (5833): 1905–1907.
65. Ohtsuki H, Iwasa Y, Nowak MA (2009) Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* 457 (7225): 79–82.
66. Sigmund K, De Silva H, Traulsen A, Hauert C (2010) Social learning promotes institutions for governing the commons. *Nature* 466 (7308): 861–863.
67. Helbing D, Szolnoki A, Perc M, Szabo G (2010) Evolutionary Establishment of Moral and Double Moral Standards through Spatial Interactions. *PLoS Comp Biol* 6: e1000758, doi:10.1371/journal.pcbi.1000758.
68. Szolnoki A, Szabo G, Czakó L (2011) Competition of individual and institutional punishments in spatial public goods games. *Phys Rev E* 84: 046106.
69. Hagen E, Hammerstein P (2006) Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology* 69 (3): 339–348.
70. West SA, El Mouden C, Gardner A (2011) Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior* 32: 231–262.
71. Binmore K, Shaked A (2010) Experimental economics: Where next? *Journal of Economic Behavior & Organization* 73 (1): 87–100.
72. Fehr E, Schmidt KM (2010) On inequity aversion: A reply to binmore and shaked. *Journal of Economic Behavior & Organization* 73 (1): 101–108.
73. Gintis H (2010) Towards a renaissance of economic theory. *Journal of Economic Behavior & Organization* 73 (1): 34–40.
74. Eckel C, Gintis H (2010) Blaming the messenger: Notes on the current state of experimental economics. *Journal of Economic Behavior & Organization* 73 (1): 109–119.