CIRANO
Allier savoir et décision

The Evolution of Morals under Indirect Reciprocity

ALEXIA **GAUDEUL**
CLAUDIA **KESER**
STEPHAN **MÜLLER**

2019S-29
CAHIER SCIENTIFIQUE
CS

2019s-29

The Evolution of Morals under Indirect Reciprocity

Alexia Gaudeul, Claudia Keser, Stephan Müller

Série Scientifique
Scientific Series

Montréal
Décembre/December 2019

CIRANO
Allier savoir et décision

Centre interuniversitaire de recherche en analyse des organisations

CIRANO

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du gouvernement du Québec, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Quebec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the government of Quebec, and grants and research mandates obtained by its research teams.

Les partenaires du CIRANO

Partenaires corporatifs
Autorité des marchés financiers
Banque de développement du Canada
Banque du Canada
Banque Laurentienne
Banque Nationale du Canada
Bell Canada
BMO Groupe financier
Caisse de dépôt et placement du Québec
Canada Manuvie
Énergir
Hydro-Québec
Innovation, Sciences et Développement économique Canada
Intact Corporation Financière
Investissements PSP
Ministère de l'Économie, de la Science et de l'Innovation
Ministère des Finances du Québec
Mouvement Desjardins
Power Corporation du Canada
Rio Tinto
Ville de Montréal

Partenaires universitaires
École de technologie supérieure
École nationale d'administration publique
HEC Montréal
Institut national de la recherche scientifique
Polytechnique Montréal
Université Concordia
Université de Montréal
Université de Sherbrooke
Université du Québec
Université du Québec à Montréal
Université Laval
Université McGill

Le CIRANO collabore avec de nombreux centres et chaires de recherche universitaires dont on peut consulter la liste sur son site web.

ISSN 2292-0838 (en ligne)

The Evolution of Morals under Indirect Reciprocity

Alexia Gaudeul [*], *Claudia Keser* [†], *Stephan Müller* [‡]

Abstract/Résumé

We theoretically and experimentally study the evolution of strategies reflecting different moral judgments under indirect reciprocity. We fully characterize the evolutionary stable equilibria. In all cooperative equilibria multiple strategies coexist. This offers an explanation for the heterogeneity in moral judgments among humans. The prescribed behavior of the equilibrium strategies can rationalize the design of empirical examples of reputation systems, which are set up to resolve problems of moral hazard. In our laboratory experiment, we find that more than 75% of participants play strategies that belong to the predicted equilibrium set.

Keywords/Mots-clés: Indirect Reciprocity, Cooperation, Evolution, Experiment

JEL Codes/Codes JEL: C73, C91, D83

[*] University of Göttingen. Platz der Göttingen Sieben 5, 37073 Göttingen (Germany), *E-mail: alexia.gaudeul@wiwi.uni-goettingen.de*

[†] University of Göttingen. Platz der Göttingen Sieben 3, 37073 Göttingen (Germany), *E-mail: c.keser@uni-goettingen.de*

[‡] University of Göttingen. Corresponding author, Platz der Göttingen Sieben 3, 37073 Göttingen (Germany), *E-mail: stephan.mueller@wiwi.uni-goettingen.de*

1 Introduction

Many philosophers share the view that the function of morality is to induce interpersonal coordination and to produce mutually beneficial cooperative patterns (Prinz, 2007; Sinclair, 2012). Joyce (2006) emphasizes that the moralization of our practical lives serves our long-term interests by supplying license for punishment, and justification for likes and dislikes. From an economist's perspective, morality might therefore be of particular relevance for interactions where people have an incentive to free ride on the cooperation of others. The role of moral judgment might be even more pronounced in situations where the mechanism of direct reciprocity cannot sustain cooperation. In such adverse environments the concept of indirect reciprocity has been proposed to explain the evolution of cooperation (for a survey, see Nowak, 2006).[1] Contrary to direct reciprocity, under indirect reciprocity (Trivers, 1971; Alexander, 1987) a cooperative act is not reciprocated by the receiver of that act but by a third party. Promoting cooperation through this mechanism requires individuals to carry an observable reputation which informs other members of the society about their past behavior.

Greif (1989) provides a historical example of such a reputation system. In the 11th Century a group of Mediterranean traders relied on other coalition members as agents to complete some of their business dealings overseas. The immanent moral hazard problem was solved via an informal reputation mechanism described by Greif as follows. "[A]ll coalition merchants agree never to employ an agent that cheated while operating for a coalition member. Furthermore, if an agent who was caught cheating operates as a merchant, coalition agents who cheated in their dealings with him will not be considered by other coalition members to have cheated." Thus, from a moral perspective defection on someone who defected before was justifiable. The mechanism of indirect reciprocity has also been put forward as a rationale for reputation systems used in online markets such as Amazon and eBay with many one-shot interactions (Bolton et al., 2004; Resnick et al., 2006; Diekmann et al., 2014).

The economic literature has established with quite some generality that community enforcement can sustain cooperation as a social norm in random matching games played by rational forward-looking agents (Kandori, 1992; Okuno-Fujiwara and Postlewaite, 1995; Takahashi, 2010). Apart from this, and despite the theoretical appeal of indirect reciprocity and its potential importance for the initiation of cooperation among strangers,

[1] In this paper we focus on what has been called *downstream* reciprocity, i.e., you are helped by somebody because you helped someone else before. Nowak and Roch (2006) show that so called *upstream* reciprocity, i.e., you help somebody because somebody else has helped you, cannot lead to the evolution of cooperation.

there is surprisingly little research in economics on this topic. Notable exceptions are the recent theoretical contributions by Berger (2011), Berger and Grüne (2016), and the experimental research by, for instance, Bolton et al. (2005), Seinen and Schram (2006), Engelmann and Fischbacher (2009), and Charness et al. (2011). Most research on indirect reciprocity has been conducted in the field of evolutionary biology. This literature is in the tradition of the seminal paper of Nowak and Sigmund (1998) and focuses on the identification of specific reputation mechanisms and its informational requirements for cooperation to evolve.[2] In that research it is assumed that all members of a society obey the same reputation mechanism, i.e., all individuals share the same notion of what is considered to be good and what is bad behavior.[3]

One main limitation of this approach is that it precludes the apparent heterogeneity of what agents consider to be good or bad behavior which in turn determines their inclination to help others. This is because of the analytic intractability of the extension of the model to multiple reputation mechanisms simultaneously at work. Apart from its intractability, the coexistence of different reputation mechanisms would make it difficult to provide a plausible story of information processing as each agent is required to carry multiple labels which constantly need to be collectively updated. This is because what some members of the society might consider to be good behavior might be bad from the perspective of others. Consequently, the apparent question regarding the coevolution of different reputation mechanisms reflecting different moral judgments cannot be addressed. As another limitation we consider the difficulties of generating testable hypotheses because in this model strategies condition on individuals' reputation which is not directly observable in real-life.

In this paper we overcome these shortcomings and provide an analytically tractable model of indirect reciprocity which allows the study of the evolution of different inherited strategies reflecting different moral judgments. We obtain this by discarding the concept of a single, universally shared reputation mechanism. In our model behavior may simply condition on information about past behavior which is also observable by the econometrician or that is even perfectly controllable by the experimenter. This approach also bridges the gap between most previous theoretical research and laboratory studies on indirect reciprocity. This is because in implementations in the lab there is no commonly shared reputation mechanism but participants are simply provided with publicly available (higher-order) information about past play of other players. We test our theoretical

[2]The most comprehensive study in this regard is Ohtsuki and Iwasa (2006).

[3]The only exception is a recent paper by Yamamoto et al. (2017) who study the coevolution of reputation mechanisms in an agent-based model.

predictions in a laboratory experiment.

Our contribution is threefold. First, we contribute to the theoretical literature on indirect reciprocity. We fully characterize the set of all evolutionary stable equilibria for the considered class of strategies. Quite surprisingly, there are only two stable cooperative equilibria in the 15-dimensional population state space. Both are characterized by the coexistence of different morals and the presence of players who's behavior depends on so-called second-order information. The two cooperative equilibria, (1) a single population state with two strategies, and (2) a linear equilibrium set with two additional strategies, emphasize the role of a particular second-order strategy. The implicit moral of that strategy matches with the real-life societal judgment as in the historical example of Greif (1989). That is, not cooperating with someone who cooperated before is a particular reprehensible behavior which ought to be punished. Hoever, not cooperating with someone who himself did not cooperate before, is regarded as a justifiable defection and therefore not punished. Our results also add to the ongoing discussion about the minimal informational requirements for community enforcement to enable cooperation. In a recent and very rich paper Heller and Mohlin (2017) show, among others, that in a related game a single observation about the opponent's past behavior may sustain cooperation. In their setting the observation is drawn randomly from the entire history of the partner. Thus, "[...] memory of past interactions is assumed to be long and accurate but dispersed." In our setting, community enforcement relies on information of a higher order but only the last informational update needs to be remembered.

Second, we add new insights to the experimental literature on indirect reciprocity or, more general, to experiments on the repeated prisoner's dilemma with stranger matching. We conducted a laboratory experiment to test (i) whether elicited strategies correspond to the equilibrium strategies predicted by our theory, and (ii) whether comparisons across treatments correspond to the theoretically predicted comparative statics with respect to the cost-benefit ratio (CBR) of cooperation. We run two treatments which differ in their CBRs. It turns out that more than 75% of the participants use one of the four strategies predicted by our theory. Moreover, differences in the composition of strategies across treatments are in line with the theory's comparative statics.

Third, we contribute to the recent economic literature which explores the evolutionary rationales for the existence of moral judgments as important drivers for individual decision-making (e.g., Bergstrom, 2009; Alger and Weibull, 2013; Alger and Weibull, 2016). Alger and Weibull (2013) show in a very general setting that when individuals' preferences are their private information, a particular convex combination of selfishness

and morality stands out as being evolutionarily stable. This combination, called *homo hamiltonensis*, which mirrors the degree of assortativity, is the unique evolutionary stable type. The authors mention that "[t]he uniqueness hypothesis is made for technical reasons, and it seems that it could be relaxed somewhat, but at a high price in analytical complexity." In this regard, our analytically tractable model provides a rationale for the coexistence of different morals for the specific case of the indirect reciprocity setting.

2 Theory

2.1 The Model

Consider a large population of infinitely-lived individuals. For each integer round $t=1,2,...$, each player randomly finds another player and engages in the so called helping game. That is, in these pairwise encounters, one of the players is randomly assigned to the role of the mover, the other person is the receiver. The assignment of players excludes the possibility of direct reciprocity. In the pairwise game, the mover can either *keep* (defect, D) or *give* (cooperate, C). In the latter case, the payoff to the mover is $-c$, whereas the receiver gets b, where $b>c>0$. In the former both players receive a payoff of 0. For the sake of notational convenience, we will make the usual assumption that each player actually plays in both roles at the same time during an interaction.

There is some consensus in the literature that the provision of so-called second-order information, i.e., information not only about the partner's past behavior but also about the action of the partner's former partner, is necessary and sufficient for cooperation to evolve under indirect reciprocity (e.g., Panchanathan and Boyd, 2003; Ohtsuki and Iwasa, 2004).[4] We therefore assume that each player at time t carries (second-order) information about her chosen action in the last period $t-1$ and about the other player's action at $t-2$. Given the two different actions there are four different labels on which players can condition their behavior. The label CD, for instance, reveals that the considered player cooperated in $t-1$, while this player's partner defected at $t-2$. Thus, there are $2^4=16$ strategies assigning C or D to each of the labels. Any strategy can be identified by a quadruple in $\{0,1\}^4$, where the first/second/third/last element specifies the behavior for the label $CC/CD/DC/DD$, and 1(0) indicates $C(D)$.

Thus, for example, the moral to unconditionally help is implicitly captured by the strategy coded as $(1,1,1,1)$. A moral judgment which only refers to the past behavior of

[4]See Berger (2011) for a qualification of this claim. He shows that with a tolerant scoring rule as a reputation mechanism cooperation can evolve without higher-order information.

the current opponent and only helps if the other player helped before is captured by the first-order discriminating strategy $(1, 1, 0, 0)$. As a final example, consider the implicit judgment of the strategy $(1, 1, 0, 1)$. This strategy induces a player to always help except for the case where the current opponent defected with a formerly cooperative person. This behavior matches with the prescribed behavior of the reputation mechanism in the historical example given by Greif (1989).

Let x_s denote the share of players applying strategy s, and let $p_{s|XY}, X, Y \in \{C, D\}$ denote the probability that an individual playing strategy s cooperates when facing a player with the label XY. In the absence of perceptional or execution errors these label-contingent probabilities are either 0 or 1. For example, the strategy s which cooperates with a player showing histories CC, CD, or DC, but defects toward players with a DD-label gives rise to $p_{s|CC}{=}p_{s|CD}{=}p_{s|DC}{=}1$ and $p_{s|DD}{=}0$.

The literature on indirect reciprocity has stressed the importance of errors in action and information processing (e.g., Leimar and Hammerstein, 2001; Panchanathan and Boyd, 2003). Intuitively, if no one commits an error at all, difference between several strategies is lost, which results in many strategies being neutral with respect to payoffs. As common, we will incorporate two types of errors. First, we will consider execution errors (γ), i.e., players defect although they intended to help. We will not consider errors of the type of unintended help. Second, we will allow for perceptional errors (ϵ), i.e., players may misperceive the actual label of a player. More precisely, we will assume that any label XY can be mispercieved with equal probability $\frac{\epsilon}{3}$ like any other label. We will assume that individual errors are independent across time and individuals. Let $\mathbb{1}_s(XY) \in \{0, 1\}$ denote the indicator function which equals 1 if strategies s implies cooperation given an opponent's label XY. We then can express $p_{s|XY}$, the probability of cooperating for strategy s conditional on the receiver's label, by:

$$p_{s|XY} = \Big(\mathbb{1}_s(XY)(1-\epsilon) + \sum_{X'Y' \neq XY} \mathbb{1}_s(X'Y') \frac{\epsilon}{3} \Big)(1-\gamma) \tag{1}$$

Furthermore, let $p_{XY|s}(t), X, Y \in \{C, D\}$ denote the probability at time t that an individual playing strategy s carries the label XY. Finally, let $p_{XY}(t), X, Y \in \{C, D\}$ denote

the share of the label XY in the population at time t. We can then write $p_{XY|s}(t)$ as:

$$p_{CC|s}(t)=p_{s|CC}\cdot p_{CC}(t-1)+p_{s|CD}\cdot p_{CD}(t-1) \tag{2}$$

$$p_{CD|s}(t)=p_{s|DC}\cdot p_{DC}(t-1)+p_{s|DD}\cdot p_{DD}(t-1) \tag{3}$$

$$p_{DC|s}(t)=(1-p_{s|CC})\cdot p_{CC}(t-1)+(1-p_{s|CD})\cdot p_{CD}(t-1) \tag{4}$$

$$p_{DD|s}(t)=(1-p_{s|CC})\cdot p_{DC}(t-1)+(1-p_{s|CD})\cdot p_{DD}(t-1) \tag{5}$$

That is, for example, a player with strategy i carries a CC label if he cooperates with someone who carries a CC-label or a CD-label. Taken together with the identity:

$$p_{XY}(t)=\sum_{s\in S} x_s\cdot p_{XY|s}(t),\ X,Y\in\{C,D\} \tag{6}$$

this gives us the following recurrence for the distribution of labels in the population

$$(p_{CC}(t),p_{CD}(t),p_{DC}(t),p_{DD}(t))'=W(t)(p_{CC}(t-1),p_{CD}(t-1),p_{DC}(t-1),p_{DD}(t-1))' \tag{7}$$

, with transition matrix $W(t)=\big(w(t)\big)_{ij}$ [5]

We will assume that there are two different timescales. The adjustment of the shares x_i is assumed to operate on a continuous time scale τ and to be slow compared to the dynamics of the distribution of labels. This mirrors the usual assumption that reputations dynamics are much faster than the adjustment of strategies (e.g., Berger, 2011). In other words, we treat reputation as instantly equilibrated when deriving the payoffs which determine the dynamics on the slow timescale. Under this assumption we can solve for the equilibrium distribution of labels p_{XY} for a given distribution of strategies in the population. Analytic results for fixed population states near the evolutionary stable equilibria and simulation results for arbitrary states show, that for any initial distribution of labels after 3−4 rounds the actual distribution of labels is very close to the share given in (8) (see Appendix A.2 for details).

$$p_{CD}=\frac{1}{2+\frac{w_{12}}{1-w_{11}}+\frac{1-w_{23}}{w_{24}}}=p_{DC}\ ,\ p_{CC}=\frac{w_{12}}{1-w_{11}}p_{CD}\ ,\ p_{DD}=\frac{1-w_{23}}{w_{24}}p_{CD} \tag{8}$$

We will make use of the equilibrium values given by (8) to calculate payoffs. Plugging in (8) into (5)−(8) gives us equilibrium strategy-contingent label probabilities $p_{XY|s}$. A movers' help depends on the label of his opponent and on the mover's inclination to

[5] For details, see Appendix A.2. Note that in the setting of Bolton et al (2005) only half the labels are updated each period. This, however, only slows down convergence.

cooperate given this label. Since the other player is randomly chosen, her label follows the equilibrium distribution p_{XY}. Whether a receiver receives help depends on his label and, given his label, the inclination of all other strategies to help weighted by their share in the population of the particular strategy. Thus, the payoff for strategy s at time τ in the label equilibrium (8) is given by:

$$\Pi_s(\tau; b, c, \epsilon, \gamma) = b \cdot \sum_{s' \in S} x_{s'}(\tau) \cdot p_C(s', s) - c \cdot p_C(s) \tag{9}$$

, where $p_C(s) = p_{s|CC} \cdot p_{CC} + p_{s|CD} \cdot p_{CD} + p_{s|DC} \cdot p_{DC} + p_{s|DD} \cdot p_{DD}$, and $p_C(s', s) = p_{s'|CC} \cdot p_{CC|s} + p_{s'|CD} \cdot p_{CD|s} + p_{s'|DC} \cdot p_{DC|s} + p_{s'|DD} \cdot p_{DD|s}$, i.e. $p_C(s)$ gives the probability that someone with strategy s cooperates, and $p_C(s', s)$ corresponds to the probability that a s'-player will help a s-player.

2.2 Evolutionary Stable Equilibria

In this section we study the existence of stable equilibrium points or sets under the well-known replicator dynamics, i.e.,

$$\dot{x}_s(\tau) = \big(\Pi_s(\tau; b, c, \epsilon, \gamma) - \overline{\Pi}(\tau; b, c, \epsilon, \gamma)\big) \cdot x_s(\tau), \tag{10}$$

where $\overline{\Pi}(\tau; b, c, \epsilon, \gamma) = \sum_{s \in S} \Pi_s(\tau; b, c, \epsilon, \gamma) \cdot x_s(\tau)$ denotes the average payoff.

Since the equilibrium distribution of labels p_{XY} depends on the distribution of strategies the replicator dynamics yield a system of nonlinear differential equations. Note that the payoffs $\Pi_i(\tau; b, c, \epsilon, \gamma)$ are linear in b/c. Hence, we can normalize $b \equiv 1$ and interpret $c \in (0, 1)$ as the cost-benefit ratio without changing the set of stable equilibria. We make the assumption that the two independent errors have the same magnitude, i.e., $\eta \equiv \epsilon = \gamma$, which simplifies presentation and analysis. Thus, we analyze the dynamic system:

$$\dot{x}_s(\tau) = \big(\Pi_s(\tau; c, \eta) - \overline{\Pi}(\tau; c, \eta)\big) \cdot x_s(\tau). \tag{11}$$

Given the 16 strategies there is a huge number of possible combinations which could form an equilibrium, ranging from a homomorphic population with just one strategy present to a fully mixed population where every strategy is played. The following Lemma will simplify the search for the equilibria substantially.

In any stable equilibrium all involved strategies must earn the same payoffs otherwise the shares would adjust according to (11). Intuitively, payoff differences between

two strategies result from differences in the label-contingent behavior the two strategies prescribe but not from the conditional behavior where they agree. That is, for instance, $\Pi_{(1,s_2,s_3,s_4)}(\tau;c,\eta)-\Pi_{(0,s_2,s_3,s_4))}(\tau;c,\eta)=\Pi_{(1,s'_2,s'_3,s'_4)}(\tau;c,\eta)-\Pi_{(0,s'_2,s'_3,s'_4))}(\tau;c,\eta)$ for all $s,s'\in S$. As a consequence, any payoff difference between two strategies can be decomposed based on four basic payoff differences $\Delta\Pi_i(\tau;c,\eta), i=1,...,4$, where

$$\begin{aligned}\Delta\Pi_1(\tau;c,\eta)=(1-\gamma)\Big((1-\varepsilon)p_{CC}\sum_{\tilde{s}}x_{\tilde{s}}(\tau)p_{\tilde{s}|CC}+\ldots+\frac{\varepsilon}{3}p_{DD}\sum_{\tilde{s}}x_{\tilde{s}}(\tau)p_{\tilde{s}|DD}\Big)\\ -c\Big((1-\varepsilon)p_{CC}+\frac{\varepsilon}{3}p_{CD}+\frac{\varepsilon}{3}p_{DC}+\frac{\varepsilon}{3}p_{DD}\Big)\end{aligned} \tag{12}$$

$\Delta\Pi_2(\tau;c,\eta), \Delta\Pi_3(\tau;c,\eta), \Delta\Pi_4(\tau;c,\eta)$ are defined analogously. In other words, each of the four basic payoff differences correspond to the payoff difference of two strategies which differ in the prescribed action for exactly one of the four labels, CC, CD, DC, and DD. The following Lemma provides the formal statement.

Lemma *Let* $s,s'\in\{0,1\}^4$ *then* $\Pi_s(\tau;c,\eta)-\Pi_{s'}(\tau;c,\eta)=\sum_{i=1}^{4}(s_i-s'_i)\Delta\Pi_i(\tau;c,\eta)$

Proof. All proofs are given in Appendix A.1. □

Note that a payoff difference of zero between two strategies does not *a priori* imply that the involved basic differences are also zero. Stability, however, indeed causes this implication. Intuitively, if two strategies earn the same payoff but two or more involved basic differences are not zero then there exists a mutant strategy which could successfully invade this population. This mutant strategy adopts the advantageous aspect and drops the disadvantageous one. Thus, in any stable equilibrium the basic differences for the labels where the equilibrium strategies disagree on the prescribed behavior must be zero.

Given the four basic differences there are five ($k=0,1,2,3,4$) *categories* of equilibria ranging from none of the basic differences being zero to all being zero. Within each of the categories there are $\binom{4}{k}$ *classes* of equilibria which correspond to the number of possible combinations among the four basic differences. Finally, each class contains 2^{4-k} subclasses of equilibria which correspond to the number of possible assignments (C or D) for each of the labels with a non-zero basic payoff difference. For instance, for the category with two basic differences being zero ($k=2$) there are $\binom{4}{2}=6$ classes and $2^2=4$ subclasses (see Table 1). In total, this gives us 81 subclasses. Each of theses subclasses may contain at least one stable equilibrium where multiplicity may result from non-linearities in the payoffs. However, under the restriction for the equilibria to be well-defined, i.e., to be elements

of Δ_{15}, Table 1 informs us that there are indeed only 30 potentially stable equilibria (see supplementary material for details).

category	class	Equilibria in sub-classes
0	$\Delta\Pi_i(\tau;b,\eta)\neq 0,\forall i$	$\{x_s \mid s\in S\}$
1	$\Delta\Pi_1(\tau;c,\eta)=0;\Delta\Pi_i(\tau;c,\eta)\neq 0,\forall i\neq 1$	$(x_{1,1,0,1},x_{0,1,0,1})$ $(x_{1,0,0,0},x_{0,0,0,0})$
	$\Delta\Pi_2(\tau;c,\eta)=0;\Delta\Pi_i(\tau;c,\eta)\neq 0,\forall i\neq 2$	$(x_{1,1,0,1},x_{1,0,0,1})$ $(x_{0,1,0,0},x_{0,0,0,0})$
	$\Delta\Pi_3(\tau;c,\eta)=0;\Delta\Pi_i(\tau;c,\eta)\neq 0,\forall i\neq 3$	$(x_{1,1,1,1},x_{1,1,0,1})$ $(x_{1,1,1,0},x_{1,1,0,0})$ $(x_{1,0,1,0},x_{1,0,0,0})$ $(x_{0,1,1,0},x_{0,1,0,0})$
	$\Delta\Pi_4(\tau;c,\eta)=0;\Delta\Pi_i(\tau;c,\eta)\neq 0,\forall i\neq 4$	$(x_{1,1,1,1},x_{1,1,1,0})$
2	$\Delta\Pi_i(\tau;c,\eta)=0,i=1,2;\Delta\Pi_i(\tau;c,\eta)\neq 0,i=3,4$	$(x_{1,1,1,0},x_{1,0,1,0},x_{0,1,1,0},x_{0,0,1,0})$ $(x_{1,1,0,1},x_{0,1,0,1},x_{1,0,0,1},x_{0,0,0,1})$
	$\Delta\Pi_i(\tau;c,\eta)=0,i=1,3;\Delta\Pi_i(\tau;c,\eta)\neq 0,i=2,4$	none
	$\Delta\Pi_i(\tau;c,\eta)=0,i=1,4;\Delta\Pi_i(\tau;c,\eta)\neq 0,i=2,3$	none
	$\Delta\Pi_i(\tau;c,\eta)=0,i=2,3;\Delta\Pi_i(\tau;c,\eta)\neq 0,i=1,4$	none
	$\Delta\Pi_i(\tau;c,\eta)=0,i=2,4;\Delta\Pi_i(\tau;c,\eta)\neq 0,i=1,3$	none
	$\Delta\Pi_i(\tau;c,\eta)=0,i=3,4;\Delta\Pi_i(\tau;c,\eta)\neq 0,i=1,2$	$(x_{1,1,1,1},x_{1,1,1,0},x_{1,1,0,1},x_{1,1,0,0})$ $(x_{0,1,1,1},x_{0,1,1,0},x_{0,1,0,1},x_{0,1,0,0})$
3	$\Delta\Pi_i(\tau;c,\eta)=0,i=1,2,3;\Delta\Pi_4(\tau;c,\eta)\neq 0$	none
	$\Delta\Pi_i(\tau;b,\eta)=0,i=1,2,4;\Delta\Pi_3(\tau;c,\eta)\neq 0$	none
	$\Delta\Pi_i(\tau;b,\eta)=0,i=1,3,4;\Delta\Pi_2(\tau;c,\eta)\neq 0$	none
	$\Delta\Pi_i(\tau;b,\eta)=0,i=2,3,4;\Delta\Pi_1(\tau;c,\eta)\neq 0$	none
4	$\Delta\Pi_i(\tau;c,\eta)=0,\forall i$	$(x_{1,1,1,1},...,x_{0,0,0,0})$

Table 1: Potentially stable equilibria.

It turns out that only four of the remaining 30 equilibria can be stable since for all other cases there exists a non-equilibrium strategy which earns strictly higher payoffs. Thus, such equilibria can succesfully be invaded and are therefore not stable. Among the four remaining candidates are a homomorphic population with $x_{0,0,0,0}=1$ and an equilibrium with $x_{1,1,1,1}+x_{1,1,0,1}=1$. The former corresponds to a population state where everybody unconditionally defects. The latter is characterized by the coexistence of unconditional cooperators and conditional cooperators who intentionally defect if and only if the label DC is observed. That is, if they encounter a receiver who did not help someone who helped in the previous period. Thus, this strategy makes use of second-order information.

Moreover, there is the equilibrium constituted by the strategies $(1,1,1,1)$,$(1,1,1,0)$, $(1,1,0,1)$, and $(1,1,0,0)$ which is indeed a linear equilibrium set which is based on three equations $\Delta\Pi_i(t;c,\eta)=0, i=3,4$, and $x_{1,1,1,1}+x_{1,1,1,0}+x_{1,1,0,1}+x_{1,1,0,0}=1$ and four unknown.

The first of the two additional strategies $(1,1,1,0)$ prescribes defection if and only if the label DD is observed by the mover. The second strategy only makes use of first-order information as it induces defection whenever the receiver previously defected. Finally, there is an equilibrium where all four basic differences vanish. Again, this is an equilibrium set which may contain all 16 strategies, i.e., fully mixed population states.

We are interested in two things. First, we want to determine whether the respective equilibrium is stable. Second, in the case of stability we want to derive the set of parameters $(\frac{c}{b}, \eta)$ such that the equilibrium under consideration is stable. We consider stability first and then turn to the set of parameters. The equilibrium $x_{0,0,0,0}=1$ is stable since in this population state the strategy $(0,0,0,0)$ earns strictly higher payoffs than any other strategy. The equilibrium which involves the two strategies $x_{1,1,1,1}$ and $x_{1,1,0,1}$ is stable if and only if the non-zero basic payoff differences favor these strategies, i.e., $\Delta\Pi_i(\tau; c, \eta)>0, i\neq 3$. Sufficiency results from the fact that an increase in the share $x_{1,1,1,1}$ decreases the profit for strategy $(1,1,1,1)$ relative to $(1,1,0,1)$. The conditions $\Delta\Pi_i(\tau; c, \eta)>0, i\neq 3$ imply that any strategy which deviates from the behavior where both equilibrium strategies agree (labels CC, CD, and DD) will earn a strictly lower payoff. Taken together, this ensures that any small perturbations will eventually vanish and equilibrium will be restored.

The one-dimensional equilibrium set with strategies $(1,1,1,1), (1,1,1,0), (1,1,0,1)$, and $(1,1,0,0)$ is stable if and only if the non-zero basic payoff differences favor these strategies, i.e., $\Delta\Pi_i(\tau; c, \eta)>0, i=1,2$. The sufficiency of this condition results from the fact that under this condition all non-zero eigenvalues of the matrix of the associated linear system have strictly negative real parts. Furthermore, the unique eigenvalue of zero is associated with the eigenvector spanning the equilibrium set. The second set is a 12-dimensional subset of Δ_{15} which makes it analytically intractable. However, we can show that the equilibrium set contains an unstable region such that a neutral drift will eventually shift the population state into that region. Thus, this equilibrium is unstable.

Taken together, our evolutionary analysis revealed a remarkably small set of only two stable cooperative equilibria. Moreover, these equilibria have a simple structure, involving only a small set of strategies which, as we see below, cannot coexist, i.e., for each CBR allowing cooperation to emerge there is a unique cooperative equilibrium. The following Proposition summarizes these insights.

Proposition *In the helping game there exist three evolutionary stable equibria:*

E_1: $\{x\in\Delta_{15}\big|x_{(0,0,0,0)}=1\}$ *is stable for all* $\frac{c}{b}\in(0,1)$.

E_2: $\{x\in\Delta_{15}\big|x_{(1,1,1,1)}+x_{(1,1,0,1)}=1,\ x_{(1,1,1,1)}=1-\frac{3(1-c)}{2\eta}+\frac{3c}{2\eta}\sqrt{1+\frac{1}{c^2}+\frac{2\eta(2+\eta(3-4\eta))-6}{c(1-\eta)^2(3-4\eta)}}\}$ *is stable if and only if* $\frac{c}{b}\in\big(0,\frac{3-10\eta+8\eta^2}{15-22\eta+8\eta^2}\big)$.

E_3: $\{x\in\Delta_{15}\big|\sum_{s\in S_1}x_s=1,\ x_{(1,1,1,0)}=\alpha-x_{(1,1,1,1)},\ x_{(1,1,0,1)}=\beta-x_{(1,1,1,1)}\}$ *is stable if and only if* $\frac{c}{b}\in\big(\frac{3-10\eta+8\eta^2}{15-22\eta+8\eta^2},\frac{3-10\eta+8\eta^2}{6-4\eta}\big)$, *and* $\frac{\eta(1+2\eta)}{3-5\eta+2\eta^2}<x_{(1,1,1,1)}<1+\frac{4}{3-4\eta}-\frac{8}{3-2\eta}+\frac{3c(5-4\eta)}{(3-4\eta)^2(1-\eta)}$,

where $\alpha\equiv1+\frac{4}{3-4\eta}-\frac{8}{3-2\eta}+\frac{3(5-4\eta)c}{(3-4\eta)^2(1-\eta)}$, $\beta\equiv1+\frac{3}{1-\eta}-\frac{4}{3-4\eta}-\frac{4}{3-2\eta}-\frac{3(5-4\eta)c}{(3-4\eta)^2(1-\eta)}$, *and* $S_1=\{x_{(1,1,1,1)},x_{(1,1,1,0)},x_{(1,1,0,1)},x_{(1,1,0,0)}\}$.

The following corollary characterizes the equilibria of the proposition in terms of the prescribed behavior and comparative statics. The comparative statics for equilibrium E_2 are derived via the derivatives of the equilibrium population state with respect to the model parameters. For the linear equilibrium set E_3 we can analyze the comparative statics by focusing on the boundaries of this set. It turns out that if the cost-benefit ratio increases the equilibrium set is shifted such that the share of the second-order discriminators playing $(1,1,0,1)$ decreases whereas all other shares increase (see Figure 2).[6] The same comparative statics result for the error η.

Corollary

(1) In E_1 everybody always defects, irrespective of the opponent's label.

(2) In E_2 some players are unconditional cooperators and some defect if and only if the receiver's label is DC. The share of unconditional cooperators decreases in the cost-benefit ratio c/b and the error η.

(3) In E_3 all players cooperate in the case of label CC or CD. In the case of an opponent with labels DC or DD some cooperate, others defect. The shares of unconditional cooperators, first-order discriminators increase and $(1,1,1,0)$-second-order discriminators increase whereas the share of $(1,1,0,1)$-second-order discriminators decreases in the cost-benefit ratio c/b and the error η.

[6]For this equilibrium set the lower (upper) bound for the share of the unconditional cooperators is given by $\frac{\eta(1+2\eta)}{3-5\eta+2\eta^2}\left(1+\frac{4}{3-4\eta}-\frac{8}{3-2\eta}+\frac{3c(5-4\eta)}{(3-4\eta)^2(1-\eta)}\right)$. Hence, the $x_{(1,1,1,1)}$-coordinate of the boundaries of equilibrium set E_3 increases in c and η.

The intuition behind our results is the following. Under the smallest errors of perception and implementation unconditional cooperation uniquely maximizes the number of cooperative acts and is therefore the efficient behavior. A homomorphic population of unconditional cooperators is, however, vulnerable to the invasion of defectors who have a relative advantage if cooperation is non-discriminatory. Importantly, the defectors will primarily receive the label DC as most of the time they interact with formerly cooperative players. Thus, in order to discipline such defectors and prevent an invasion unconditional cooperators need to be safeguarded by discriminating cooperators. These players are indeed second-order discriminators who refuse to help if and only if they are matched with a person with a DC reputation. The required share, $x_{(1,1,0,1)}$, increases as cooperation becomes relatively more costly, i.e., as CBR increases. This is because a higher share $x_{(1,1,0,1)}$ increases the likelihood of defectors getting punished and this compensates for the relative disadvantage of cooperators as a result of a higher CBR. Now, if the CBR further increases defection on cooperative behavior eventually turns profitable. To circumvent this, punishment needs to be intensified. Since defectors are relatively more likely to generate the DD reputation the second-order discriminator strategy $x_{(1,1,1,0)}$ becomes an equilibrium strategy to back up cooperation. As a consequence of our lemma this also introduces the first-order discriminator strategy $x_{(1,1,0,0)}$ (see Figure 1). Finally, defection when facing the label CD can never be part of a cooperative equilibrium strategy. Intuitively, defection conditional on the label CD does not punish defectors who primarily carry labels DC and DD but only the cooperators predominantly carrying labels CC and CD. Thus, such behavior cannot stabilize cooperation.

Before we analyze the composition and the comparative statics of the equilibria in more detail we turn to the set of parameters $(\frac{c}{b}, \eta)$ satisfying the necessary and sufficient conditions for stability presented in the proposition above. The parameter regions for each of the cooperative equilibria are depicted in Figure 1.

As revealed in Figure 1 equilibrium (2) is stable for low CBRs and equilibrium (3) for intermediate levels. Figure 1 also shows that the considered errors in perception and implementation limit the range of CBRs for which cooperation may be sustained. Intuitively, if the chance of unintended defection increases cooperators are also more often punished because of the presence of discriminators. As a consequence, cooperation earns less and the population state may evolve toward full defection. Each of the pairs of parameters depicted in Figure 1 corresponds to a particular equilibrium (set). Figure 2 illustrates the distribution over equilibrium strategies for each pair of parameters.

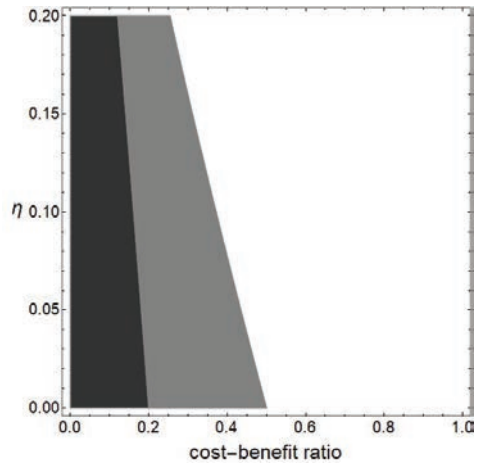

Figure 1: Region of parameter values for the equilibria of the Proposition (dark gray region – equilibrium E_2; light gray region – equilibrium E_3.

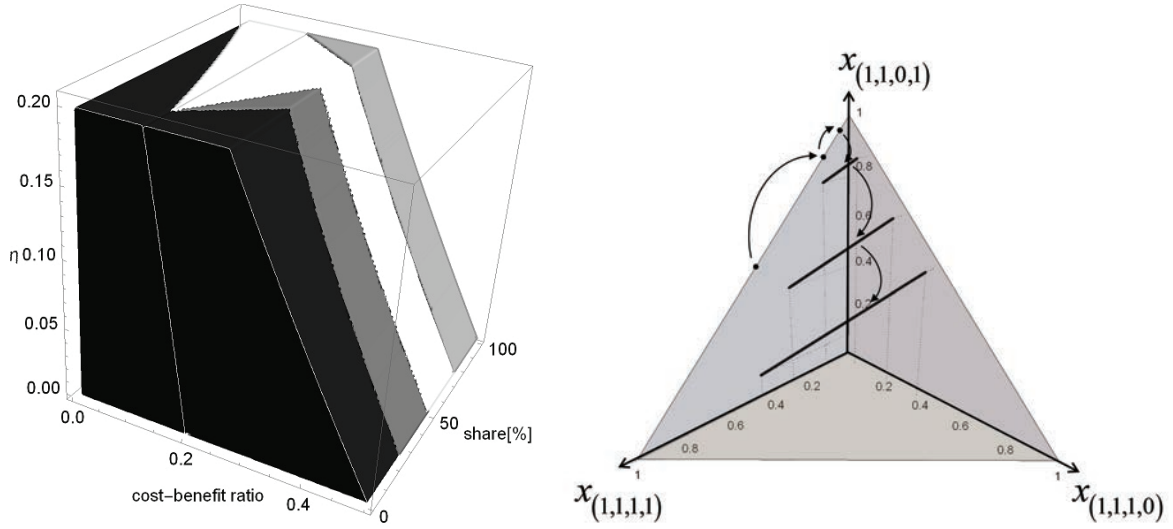

Figure 2: Left: Distribution of strategies in E_2 and E_3: black – $x_{(1,1,1,1)}$; white – $x_{(1,1,0,1)}$; light gray – $x_{(1,1,0,0)}$; dark gray – $x_{(1,1,1,0)}$. Right: Equilibria E_2 and E_3 in the 3-simplex for different cost-benefit ratios. The arrows indicate increasing cost-benefit ratios.

For the equilibrium set E_3 this requires the selection of a particular population state. In this regard, Figure 2 focuses on the center of gravity for the linear equilibrium set E_3. In line with the corollary, Figure 2 reveals that in the region associated with equilibrium E_2 the share of second-order discriminators $x_{(1,1,0,1)}$ increases in the cost-benefit ratio. When this equilibrium looses stability the minimum level of unconditional cooperators which can be sustained by this equilibrium is reached. The emergence of another second-order strategy, $(1, 1, 1, 0)$, and of the first-order discriminator $(1, 1, 0, 0)$ increases in CBR

and partially substitutes $(1, 1, 0, 1)$. Additionally, Figure 2 indicates that the distribution of equilibrium strategies is fairly robust to errors in the perception of labels and the implementation of strategies.

3 Experiment

We now turn to the design and the results of our laboratory study. In section 3.1 we present the design of the experiment. For the detailed instructions we refer the reader to appendix A.6. We derive our hypotheses in section 3.2 and present the results of the experiment in section 3.3.

3.1 Design of the experiment

Participants were assigned into groups of 12, with the group being assigned either a "low" or "high" cost of giving (c=5 ECU, 10 ECU). The benefit of giving was always b=25 ECU. Participants were matched in pairs within their own group over 11 periods. No two participants were matched together more than once, and they were made aware of this in advance (perfect stranger matching).

In each period, participants were asked to choose between keeping or giving to their partner. If a mover gave, then his payoff was 75 ECU and that of the receiver was $40 + b$ ECU. If a mover kept, then his payoff was $75 + c$ ECU and that of the receiver was 40 ECU. The exchange rate was set at 1 ECU=0.10 € and participants were additionally paid a 4 € show-up fee.

In the first period, participants were given no information about their respective partner. In the second period, they were told what their partner did last period (first-order information). In the third and subsequent periods, they were told what their partner did last period and what their partners' partner had done in the previous-to-last period (second-order information). We illustrated the game played and the meaning of the information given by representing participants on the screen as shown in Figure 3 for periods 3 to 11 (equivalent representations were shown for periods 1 and 2).

Furthermore, in order to emphasize the consequences of actions taken and help strategic thinking, participants were told, using the same representation, and between each period, what information would be given to their next partner about themselves in the following period.

Our experiment was inspired by Bolton, Katok and Ockenfels (2005) ("BKO") but differs in a number of respects. BKO told participants in each period whether they

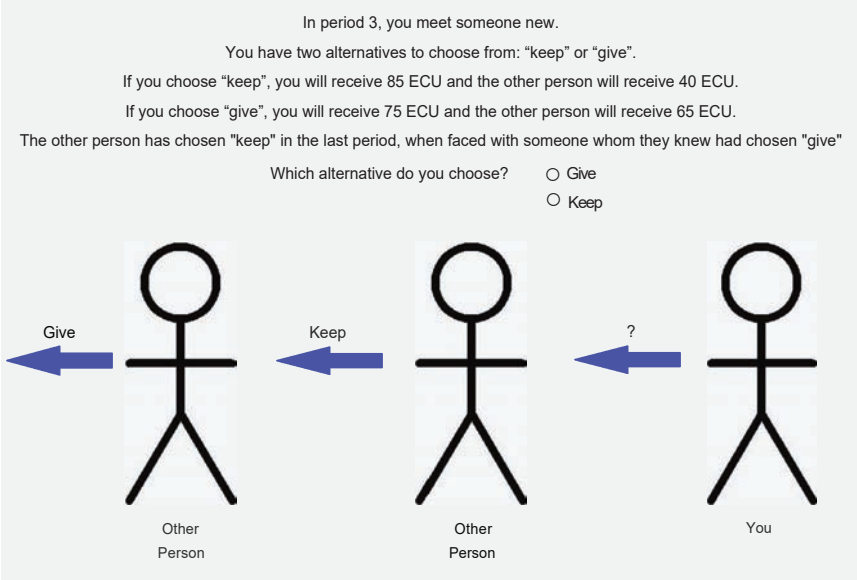

Figure 3: Visual and written representation of the game from period 3 onwards.

were movers or receivers while we assigned the role of mover or receiver ex-post: both participants of a pair in a given period had to make a decision to give or to keep, but only one of them would see their decision implemented at the payment stage. This increased the amount of data collected (11 decision periods rather than 7). In BKO participants were informed in each round about the outcome of the interaction. In our experiment participants were informed only about the outcome of the round which was selected to be payoff relevant at the end of the experiment. This excludes confounds due to upstream reciprocity which is not part of our theory. BKO told participants how many periods they would play while we did not inform participants of the number of periods in the experiment. This avoids the end-game effect, with reduced giving, that occurs in BKO.

BKO used direct elicitation, i.e., participants only made decisions for actual labels faced in the experiment, while we combined this with another elicitation method. We used the strategy method (Selten, 1967) in period 5 and period 11 in order to know what action would have been chosen for every potential reputational information. In these periods participants had to choose an action for all four possible labels they could encounter. If one of those periods was chosen for payment then participants were paid for the action the mover took with respect to the actual label of the receiver. This solves the issue that we do not know what strategy was played by many participants in BKO because most of them were not faced with the full range of reputational information. We use the strategy method only in two periods, while other periods are designed to give

participants experience in playing the game, and thus inform their answers under the strategy method.

Finally, BKO chose cost-benefit ratios c/b=1/5 and c/b=3/5 while we use ratios c/b=1/5 and c/b=2/5. We have chosen these values for several reasons. First, our model predicts $x_{0,0,0,0}$=1, that is, no giving, as the unique stable equilibrium for $c/b \geq 1/2$. The prediction for c/b=3/5 is therefore straighforward and of little interest for us. Indeed, we are mainly interested in CBRs that are consistent with the emergence of cooperative equilibria, and we focus on the type of strategies that sustain such equilibria. We chose c/b=1/5 to allow a direct comparison with the results in BKO. For c/b=1/5 our model predicts E_3 to be the unique stable equilibrium. We will analyze whether observed strategies correspond to the predicted set of equilibrium strategies. We further test the comparative statics in our model by choosing another CBR, c/b=2/5, that has the same equilibrium set prediction in terms of strategies employed, but differs in terms of the proportions of predicted strategies in the population.

Procedure

We ran our experiment in March 2017 at the Göttingen Laboratory of Behavioral Economics (GLOBE) in Göttingen. The experiment was computerized using the Zurich Toolbox for Ready-made Economic Experiments (z-Tree; Fischbacher, 2007) and participants were recruited using the Online Recruitment System for Economic Experiments (ORSEE; Greiner, 2015). In accordance with the Declaration of Helsinki, all participants were requested to read an online consent form and agree with its terms (by clicking) before registering to take part in an experiment.[7] Participants were guaranteed the anonymity of the data generated during the experiment.[8]

We ran seven experimental sessions with 24 participants each, for a total of 168 participants. Each session had two groups of 12, one for each treatment. In total, we observed 14 groups of 12 participants, that is, seven groups for each of the cost treatments. Participants ranged in age from 18 to 51, with a mean of 25. 44% were men, 94% German and 95% students, of which 45% studied economics. The experimental sessions lasted approximately 90 minutes and participants earned an average of 10.70€ (min = 8€, max = 12.50€). This payoff includes a 4€ base payment for showing up, as well as a payment proportional to individual payoff in one randomly selected period of the experiment.

[7]Rules of the GLOBE are available at http://www.lab.wiwi.uni-goettingen.de/public/rules.php

[8]The GLOBE's privacy policy is available at http://www.lab.wiwi.uni-goettingen.de/public/privacy.php

Payment We implemented some aspects of the PRINCE procedure (Johnson et al., 2015) by selecting in advance of each session the period and role that would be paid for each participant, and giving them this information on a piece of paper in a sealed envelope at the beginning of the experiment, to be opened only at its end.[9] The period was drawn randomly, as was the role of each member of the pairs formed in that period. The sealed envelopes were labeled with a cabin number. Participants drew an envelope at random when entering the laboratory and went to the corresponding cabin. They were instructed not to open their envelope until we gave them the signal to do so. At the end of the experiment, and after we checked that each envelope was still closed, we gave participants the signal to open them. Participants then had to input the information contained in their envelope on their computer. We checked this was done correctly before letting them proceed to the payment stage. At that stage, along with their payoff, movers were reminded of the action they had taken in the payment period and of the information that was available to them about their counterpart in that period. Receivers were informed in a similar way.

The advantage of this procedure was to emphasize that the assignment to a role and the choice of a payment period are not affected by a participant's behavior during the experiment. It also emphasizes that any period may be relevant, and that the role played in that period is not known. Finally, this procedure makes it easy to explain to participants in what role and in what period they will be paid: we inform them of this on the paper inside their closed envelope.

Questionnaire The experiment was followed by a socio-economic questionnaire, a modified cognitive-reflection test (Frederick, 2005), and questions about the participant's experience and perception of the experiment. Translated instructions and a list of questions are shown in Appendix A.6.

3.2 Hypotheses

Our theoretical results lead to two testable predictions. Equilibrium set E_3 is the unique stable equilibrium for the cost-benefit ratios in both of our treatments (see Figure 1). We first test that the strategies used by participants correspond to those predicted by E_3:

Hypothesis 1 *In both treatments participants will employ one of the following strategies:* $(1,1,1,1)$*,* $(1,1,1,0)$*,* $(1,1,0,1)$*, or* $(1,1,0,0)$*.*

[9] The paper in the envelope told them "You are a [mover/recipient]. The period that will be paid is period [number]."

This hypothesis means in particular that some participants will take account of second-order information in their choices, i.e., they will adopt different behavior when faced with labels DC and DD. We then test comparative statics within E_3 based on the Corollary (see Figure 2), by comparing the frequency of strategies across treatments.

Hypothesis 2 *The shares of unconditional cooperators* $(1,1,1,1)$*, first-order discriminators* $(1,1,0,0)$*, and* $(1,1,1,0)$*-players will be higher in the high cost treatment than in the low cost treatment, while the share of* $(1,1,0,1)$*-players will be lower.*

3.3 Results

The structure of the presentation of the experimental results is as follows. First, we present summary statistics on directly elicited behavior and compare those with findings in BKO. In a second step we show the consistency of direct elicitation and the strategies elicited in periods 5 and 11. Finally, we test our two hypotheses with data on the elicited strategies.

3.3.1 Direct elicitation

Table 2 shows results from the direct elicitation method in terms of the percentage of giving depending on the label of the receiver. There was no label in the first period, labels were either C or D in the second period, and were either CC, CD, DC or DD in the third and later periods. We also show results from BKO for comparison, based on our own calculations with data provided by the authors.

Table 2 highlights that irrespective of the receiver's reputation, the average giving rates are generally higher in the low cost treatment than in the high cost treatment. We note that low overall cooperation rates in the high cost treatment c/b=3/5 of BKO are consistent with the predictions of our model, whereby cooperation breaks down at such high cost levels. The giving rates in the low cost treatment when faced with label CC are consistent with those in BKO's low cost treatment, which used the same cost-benefit ratio. Other giving rates are difficult to meaningfully compare given the low number of observations in BKO.

We observe as in BKO that label CC is dominant in the low cost treatment (66% of observations in our experiment vs. 84% of observations in BKO). As in BKO, the increase in the cost-benefit ratio leads to a wider variety of labels being found in the population, whereby label CC accounts for only 53% of observations in our high cost treatment (c/b=2/5), and 29% of observations in BKO's high cost treatment (c/b=3/5).

Label	Treatments				BKO (own calculations)			
	Low cost (c/b=1/5)		High cost (c/b=2/5)		Low cost (c/b=1/5)		High cost (c/b=3/5)	
	First period							
None	74%	(84)	67%	(84)	81%	(16)	63%	(16)
	Second period							
C	90%	(62)	79%	(56)	85%	(13)	70%	(10)
D	68%	(22)	57%	(28)	33%	(3)	33%	(6)
Average	85%	(84)	71%	(84)	75%	(16)	56%	(16)
	Periods 3–4 and 6–10							
CC	86%	(390)	82%	(312)	91%	(162)	66%	(56)
CD	85%	(68)	69%	(80)	77%	(13)	50%	(30)
DC	47%	(60)	47%	(72)	50%	(10)	26%	(35)
DD	49%	(70)	40%	(124)	71%	(7)	31%	(71)
Average	78%	(588)	67%	(588)	88%	(192)	43%	(192)
Overall average	78%	(756)	68%	(756)	86%	(224)	46%	(224)

Table 2: Giving by treatment and label, direct elicitation. Number of observations in parentheses.

With respect to the dynamics of the label distribution we observe a fast convergence to a stable distribution of labels in both treatments which warrants the corresponding assumption in our theoretical analysis (see Figure 4, Appendix A.3).

3.3.2 Strategies vs. direct elicitation

From a review of literature comparing the two methods (Brandts and Charness, 2011), we expect broad agreement in the statistics elicited under the strategy method and the direct elicitation method in our experiment. There is no meaningful test of consistency of methods on an individual level. This is because initial learning and the limitations in both the number of individual observations and in the variation in receivers' labels. We therefore assess differences between elicitation methods by comparing aggregate rates of giving by label under the strategy method (Table 3) with those elicited using the direct elicitation method (Table 2).

We find that giving rates as a function of labels are not significantly different between the strategy method and direct elicitation. The largest difference is in giving rates when faced with label DC in the high cost treatment, which are 32% under the strategy method in period 5 while they are 47% under direct elicitation. This difference is, however, not significant (two-sample test of proportions, p=0.055). Hence, aggregate giving rates indicate that behavior is consistent across elicitation methods in both treatments and

	Period 5		Period 11	
Label	Low cost (c/b=1/5)	High cost (c/b=2/5)	Low cost (c/b=1/5)	High cost (c/b=2/5)
CC	87%	81%	92%	77%
CD	88%	80%	82%	77%
DC	38%	32%	43%	44%
DD	39%	32%	42%	38%

Table 3: Giving by treatment and label, strategy elicitation (N=84 in each cell).

elicitation periods. As outlined in the design section, the main purpose of also using the direct strategy method was to give participants the opportunity to learn the game and learn their label-contingent inclination to help, which in turn allows them to formulate strategies. When turning to the tests of our hypothesis, in the following we focus on the data elicited via the strategy method.

3.3.3 Strategy profiles

We now consider the strategy profiles of our participants, which are not available in BKO since BKO used only the direct elicitation method and many participants were never faced with the full range of labels. We use strategies elicited using the strategy method in the 5th period, which occurs after two periods of experience dealing with second-order information, and in the 11th period, which is the (unannounced) last period.

Composition of strategies (Hypothesis 1) Table 4 reports the number of participants employing each type of strategy, with strategies labeled as in the theoretical part. We find that the majority of participants follow one of the four predicted equilibrium strategies. Indeed, predicted strategies account for 85% (low cost) and 76% (high cost) of individual strategies in period 5, and account for 81% (low cost) and 73% (high cost) of individual strategies in period 11. We can therefore state the following result.

Result 1 *In both treatments more than* 3/4 *of the participants follow one of the predicted equilibrium strategies.*

We consider next whether this 75% share is consistent with our hypothesis according to which 100% of participants play equilibrium strategies. In order to do so, we have to take into account individual errors in implementation that capture unintended defection in our model (parameter η). Such errors may lead participants to specify off-equilibrium strategies. We thus compute the *aggregate share of the equilibrium strategies* (ASES) we should actually expect to observe given errors of implementation if we assume that all

participants intended to play equilibrium strategies. The ASES is $(1-\eta)^2(1+\eta)^2-(\alpha+\beta)(1-\eta)^2(1+\eta)2\eta+(1-\eta)^2 4\eta^2 x_{(1,1,1,1)}$, where α and β are defined in the proposition presented in Section 2.2. This proposition also provides upper and lower bounds on $x_{(1,1,1,1)}$, which we can then translate into corresponding bounds on ASES for a given η.

	Period 5		Period 11		Stability		Stability*	
Strategy	LC	HC	LC	HC	LC	HC	LC	HC
$(\mathbf{1,1,1,1})$	**21**	**20**	**23**	**25**	**18/21**	**19/20**	**21/21**	**20/20**
$(\mathbf{1,1,1,0})$	**10**	**5**	**13**	**6**	**7/10**	**2/5**	**10/10**	**4/5**
$(\mathbf{1,1,0,1})$	**10**	**6**	**11**	**5**	**8/10**	**4/6**	**10/10**	**6/6**
$(\mathbf{1,1,0,0})$	**30**	**33**	**21**	**25**	**17/30**	**21/33**	**25/30**	**29/33**
Total in E_3	**71**	**64**	**68**	**61**	**50/71**	**46/64**	**66/71**	**59/64**
$(1,0,1,0)$	0	1	0	2	0/0	0/1	0/0	0/1
$(1,0,0,1)$	0	0	1	0	0/0	0/0	0/0	0/0
$(1,0,0,0)$	2	3	8	2	2/2	0/3	0/2	1/3
$(0,1,1,0)$	0	0	0	1	0/0	0/0	0/0	0/0
$(0,1,0,1)$	1	0	0	0	0/1	0/0	1/1	0/0
$(0,1,0,0)$	2	3	1	3	0/2	0/3	0/2	0/3
$(0,0,1,1)$	1	1	0	2	0/1	0/1	1/1	0/1
$(0,0,1,0)$	0	0	0	1	0/0	0/0	0/0	0/0
$(0,0,0,0)$	7	12	6	12	5/7	10/12	0/7	1/12
Total not in E_3	13	20	16	23	7/13	10/20	2/13	2/20
Total	84	84	84	84	57/84	56/84	68/84	61/84

LC – Low Cost (c/b=1/5); HC – High Cost (c/b=2/5). Stability*: Maintain strategy within the equilibrium set or switch to a strategy within the equilibrium strategy set.

Table 4: Distribution and stability of strategies across treatments.

We estimate individual implementation error η by comparing answers by our participants in periods 5 and 11 to what they did in periods 4 and 10. Participants received no new information about the distribution of labels between periods 4 and 5, and between periods 10 and 11, since they received no second-order information about their randomly assigned partner in periods 5 and 11. They therefore had the same information in periods 5 and 11 as in periods 4 and 10, respectively, so that differences in the decision to help in period 4 (10) for a given observed label and the decision in period 5 (11) for that label provides a good measure of potential implementation errors. The average error rate, η, was 7.5% when aggregating over treatments and periods (Table A.4 in the Appendix). The 95% exact (Clopper-Pearson) binomial confidence interval for this statistic is [4.5%, 11.5%].

Given the above confidence interval estimate for η, we obtain an average ASES of about 71% with a confidence interval of $[55\%, 82\%]$ for both treatments. Hence, the share of participants who are observed playing equilibrium strategies in our experiment is

consistent with our measure of implementation errors and the hypothesis that every player adheres to one of the predicted equilibrium strategies. Note that while individual errors of implementation can account for the infrequent observation of most off-equilibrium strategies, the observed share of unconditional defectors is higher than what could be explained with implementation errors alone. We shall elaborate on this in our discussion below.

Comparison across treatments (Hypothesis 2) We now compare the frequency of different types of strategies across treatments. We can reject the hypothesis that the frequency of use of strategies does not differ across treatments. This holds for both, period 5 (Pearson's χ^2=10.44, p=0.034) and period 11 (Pearson's χ^2=18.30, p=0.001). There are therefore significant differences in the distribution of strategies across treatments.

	$p(s_L > s_H)$	
Strategy	Period 5	Period 11
$(1,1,1,1)$	57%	37%
$(1,1,1,0)$	91%	95%
$(1,1,0,1)$	84%	94%
$(1,1,0,0)$	31%	24%

Table 5: Posterior probabilities, p, that the proportion of the considered strategy in the low cost treatment, s_L, is higher than in the high cost treatment (s_H).

We perform a Bayesian estimation of a model of multinomial proportions over all strategies, so as to express posterior probabilities that the difference in the proportions of each strategy across treatments are greater than zero (Table 5). We find that the share of unconditional cooperators and first-order discriminators is likely to be equal or higher in the high cost treatment than in the low cost treatment in period 11, while the share of $(1,1,0,1)$-players is likely to be lower. This is consistent with Hypothesis 2. In contrast to our prediction, the share of $(1,1,1,0)$-players, however, is likely to be lower in the high cost treatment. Overall, we can state the following result:

Result 2 *The mix of strategies followed by participants differs across treatment. Three out of four of the differences in the frequency of strategies are consistent with Hypothesis 2.*

Stability of strategy choices We test the stability of equilibrium E_3 by considering whether there are fewer changes of strategies between periods 5 and 11 by participants who follow equilibrium strategies than by participants who follow non-equilibrium strategies.

As a first indication of overall stability, a multinomial chi-square test of frequencies cannot reject the hypothesis that there is no change in the frequencies of strategies from period 5 to period 11, either in the low cost treatment (Pearson's χ^2=5.38, p=0.251) or in the high cost treatment (Pearson's χ^2=4.32, p=0.365).[10] There are therefore only limited changes in the mix of strategies over time in both treatments.

We then analyze whether participants who adopt strategies that correspond to our predictions do so with more consistency than participants who adopt other strategies. If this is so, this allows us to argue that other strategies correspond to mistakes by participants or to an unfinished learning process. The two columns labeled "stability" in Table 4 show how many participants use the same strategy in period 11 as in period 5. Overall, about 2/3 of participants in both treatments maintained the same strategy from period 5 to period 11. We find that strategies in the equilibrium set are more robust than those outside of it (70% stability within vs. 54% stability outside in the low cost treatment; 72% stability within vs. 50% stability outside in the high cost treatment). This finding is further strengthened when considering how many participants maintain strategies within the equilibrium set or switch within such strategies from period 5 to period 11 (last two columns in Table 4). Overall, 93% of those who employ a strategy within the equilibrium set in period 5 still do so in period 11 in the low cost treatment, and 94% in the high cost treatment. The corresponding stability for non-equilibrium strategies other than $(0, 0, 0, 0)$ is about 64% in both treatments, which is substantially lower. Participants who play the non-equilibrium strategy $(0, 0, 0, 0)$, however, are also highly likely to keep doing so: 71% do so in the low cost treatment, and 83% in the high cost treatment. These results are also supported by regression analysis (see, Table 7 Appendix A.5). We summarize these insights in an additional result.

Result 3 *Participants who use predicted equilibrium strategies do so in a more consistent way than those who apply off-equilibrium strategies. Unconditional defectors, however, are also consistent in their use of their strategy.*

4 Discussion

Since the mechanism of indirect reciprocity has predominantly been studied in the biological literature, we briefly discuss our theoretical findings in light of this strand of literature. In that literature, it is usually assumed that each player has a binary reputation score, either *good* or *bad*. This reputation is observable by all other players and

[10] We assign all non-equilibrium strategies other than unconditional defection to the same category.

each player applies a behavioral strategy that prescribes the behavior for each reputation score. A large part of the literature studied the dynamics of the behavioral strategies under the assumption that the whole population shares the same social norm, i.e., there is an agreement on what is good and what is bad. Partially, this assumption is made for technical reasons. If multiple social norms coexist then a player might be considered bad by some and good by some others, which makes the analysis of the reputation dynamics almost intractable.

One of the most comprehensive studies in this literature is Ohtsuki and Iwasa (2004). They present an exhaustive analysis of all reputations dynamics that assign a binary reputation (good or bad) to each player when his action, his current reputation, and the other player's reputation are given. They identify and characterize eight reputation dynamics, which they name the "leading eight," that can maintain a high level of cooperation. As a common characteristics, they share the property that, for someone with a good reputation, helping is considered to be good behavior whereas refusing to help is thought of as bad. They also share the characteristic that refusal to help a bad person does not undermine the reputation of a good person. The "leading eight" do not, however, all agree in their evaluation of a behavior that was undertaken by a person with a bad reputation. The leading eight differ mostly in the way in which helping someone with a bad reputation is evaluated.

We refrained from making the assumption of an universally shared social norm in society and focused on the evolution of strategies that can condition on some parts of the recent history of the game. Note that a player may cooperate for different reasons. Cooperation may not only reflect some normative judgment of the other player's past behavior but it may also involve strategic reasoning regarding the reaction of other players in the future. However, under the assumption that a player's cooperative act can be interpreted as a moral judgment of the other player's past play our results can be interpreted as an analysis of the evolution of morality under indirect reciprocity. In light of this interpretation, the aforementioned commonalities of the leading eight correspond to strategies which cooperate in the case of the label CC and refuse to help in the case of DC, which is the case for all of our conditional cooperative equilibrium strategies. Interestingly, the fact that all equilibrium strategies prescribe cooperation in the case of the label CD points toward two particular reputation systems among the leading eight.[11] The first is known as the "standing strategy" first proposed by Sugden (1986). The second differs from standing by always assigning a good reputation to someone who refuses to

[11] In the model of Ohtsuki and Iwasa (2004) this would be captured by $d_{*0C}=1$.

help a bad person. Thus, based on our results we may conjecture that a coevolutionary analysis of the leading eight reputation dynamics would reveal a particular role of these two systems.

Due to the aforementioned difficulties of a coevolutionary analysis there is only one recent study on this topic. Yamamoto et al. (2017) show with agent-based simulations that after cooperation is achieved four strategies coexist. These four strategies are exactly those four strategies constituting equilibrium set E_3.[12] Thus, our theoretical results may provide the analytic foundation of their simulation results.

Finally, as highlighted by our first experimental result, in light of individual implementation errors we cannot reject the hypothesis that all players adhere to one of the predicted equilibrium strategies. Although implementation errors can account for the infrequent occurrence of most non-equilibrium strategies, they fail to rationalize the significant and stable share of unconditional defectors. This might indicate that our theory misses a relevant driver for the participants' strategic and moral considerations. Alternatively, it might simply reflect a player who did not fully grasp the structure of the game or did not pay enough attention. The latter might induce unconditional defection as the direct rewards from defection are easier to comprehend than the uncertain and indirect benefits from cooperation.

We run multinominal logit regressions to learn more about the drivers of the heterogeneity in the choice of strategies (see Appendix A.5). Explanatory variables are information collected in the ex-post questionnaire, including socio-demographics, participants' perception and understanding of the game, and a variant of the cognitive reflection test (CRT) that was introduced by Frederick (2005) (see Appendix A.6). We are particularly interested in the relation between a participant's strategy choice and their CRT score because unconditional defection might reflect participants' limited ability or willingness to understand the rules of the game, and the more sophisticated second order discrimination might require a higher ability or willingness to reflect in order to gain a better understanding of the dynamics of the game.

In our regressions, unconditional cooperation is the baseline category. We select variables by the Akaike Information Criterion in a stepwise algorithm both adding and removing terms. Our estimation results show that the likelihood to choose unconditional defection is higher for participants who feel less similar to the other participants. In contrast, the choice of strategy $(1, 1, 0, 1)$ is positively correlated with the CRT score. Those who choose non-equilibrium strategies other than $(0, 0, 0, 0)$ do so less consistently over

[12] In the notation of Yamamoto et al. (2017) these strategies are $(GGGG)$, $(GGBG)$, $(GGGB)$, and $(GGBB)$.

the two periods of elicitation and are more likely to be female and older participants (for details, please see Appendix A.5). These results indicate that participants who choose the second-order strategy which is of particular theoretical and empirical importance, i.e., $(1, 1, 0, 1)$, stand out regarding their level of cognitive reflection. Furthermore, unconditional defection may be driven more by feeling less similar to the other participants than by moral or strategic considerations.

5 Conclusion

In this paper we present an analytically tractable evolutionary model of indirect reciprocity where strategies reflect differences in moral judgment. We also provide evidence from a laboratory experiment. Instead of assuming that each societal member abides by the same reputation mechanism, we analyze the evolution of strategies that differ in how they condition on publicly available second-order information about opponents' past behavior. We fully characterize the evolutionary stable equilibria in this game and study their comparative statics with respect to the cost-benefit ratio of cooperation. Surprisingly, there exist only two stable cooperative equilibria in the 15-dimensional population state space. These equilibria are also of low complexity, the first is a population state constituted by two strategies, the second is a linear set of population states with two additional strategies. In our laboratory experiment we employed the strategy method to gain full information about participants' strategies and we implemented two treatments with different cost-benefit ratios. More than 75% of the participants' strategies correspond to one of the four predicted equilibrium strategies. Moreover, most differences in the distribution of strategies across treatments were in line with our predictions.

The theoretical results and the experimental evidence regarding the presence of strategies which rely on second-order information reemphasize the relevance of higher-order information to promote cooperation under indirect reciprocity. Both cooperative equilibria are constituted by a heteromorphic population, which offers an explanation for the omnipresent heterogeneity in moral judgments among humans (e.g., Haidt et al., 2009; Weber and Federico, 2013). We also shed some light on the issue of selection among different reputation mechanisms. We identify a particularly important strategy which is present in both equilibria and discriminates based on second-order information. This strategy only punishes an interaction partner who had behaved non-cooperatively toward a formerly cooperative participant. It prescribes cooperation, however, if the current partner's non-cooperative behavior is justifiable in the sense that his former opponent had defected

himself. Thus, our results indicate that the reputation mechanism known as "standing" first proposed by Sugden (1986) and identified as one of the "leading eight" by Ohtsuki and Iwasa (2004) is of particular importance for the evolution of cooperation under indirect reciprocity. Moreover, we provide an analytic foundation of recent simulation results by Yamamoto et al. (2017) and a rationale for recent experimental evidence (e.g., Bolton et al., 2005; Swakman et al., 2016).

Our results further encourage the recent stand on the role and the evolution of morality taken by economists as an important explanation for human coordination and cooperation (e.g., Bergstrom, 2009; Alger and Weibull, 2013; Alger and Weibull, 2016).

References

Alexander, R.D., 1987. The biology of moral systems. Aldine de Gruyter, New York.

Alger, I., Weibull, J.W., 2013. Homo moralis – Preference evolution under incomplete information and assortative matching. Econometrica 81, 2269–2302.

Alger, I., Weibull, J.W., 2016. Evolution and kantian morality. Games and Economic Behavior 98, 56–67.

Berger, U., 2011. Learning to cooperate via indirect reciprocity. Games and Economic Behavior 72, 30–37.

Berger, U., Grüne, A., 2016. On the stability of cooperation under indirect reciprocity with first-order information. Games and Economic Behavior 98, 19–33.

Bergstrom, T., 2009. Ethics, evolution, and games among neighbors. Working Paper. University of California, Santa Barbara.

Bolton, G.E., Katok, E., Ockenfels, A., 2004. How effective are electronic reputation mechanisms? an experimental investigation. Management Science 50, 1587–1602.

Bolton, G.E., Katok, E., Ockenfels, A., 2005. Cooperation among strangers with limited information about reputation. Journal of Public Economics 89, 1457–1468.

Brandts, J., Charness, G., 2011. The strategy versus the direct-response method: a first survey of experimental comparisons. Experimental Economics 14, 375–398.

Charness, G., Du, N., Yang, C.L., 2011. Trust and trustworthiness reputations in an investment game. Games and Economic Behavior 72, 361–375.

Diekmann, A., Jann, B., Przepiorka, W., Wehrli, S., 2014. Reputation formation and the evolution of cooperation in anonymous online markets. American Sociological Review 79, 65–85.

Engelmann, D., Fischbacher, U., 2009. Indirect reciprocity and strategic reputation building in an experimental helping game. Games and Economic Behavior 67, 399–407.

Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10, 171–178.

Frederick, S., 2005. Cognitive reflection and decision making. Journal of Economic Perspectives 19, 25–42.

Greif, A., 1989. Reputation and coalitions in medieval trade: Evidence on the Maghribi traders. Journal of Economic History 49, 857–882.

Greiner, B., 2015. Subject pool recruitment procedures: Organizing experiments with ORSEE. Journal of the Economic Science Association 1, 114–125.

Haidt, J., Graham, J., Joseph, C., 2009. Above and below left–right: Ideological narratives and moral foundations. Psychological Inquiry 20, 110–119.

Heller, Y., Mohlin, E., 2017. Observations on cooperation. Review of Economic Studies 85, 2253–2282.

Johnson, C.A., Baillon, A., Bleichrodt, H., Li, Z., Van Dolder, D., Wakker, P.P., 2015. Prince: An improved method for measuring incentivized preferences. SSRN Scholarly Paper ID 2504745. Social Science Research Network. Rochester, NY. URL: `https://papers.ssrn.com/abstract=2504745`.

Joyce, R., 2006. The evolution of morality. MIT press.

Kandori, M., 1992. Social norms and community enforcement. Review of Economic Studies 59, 63–80.

Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocity. Proceedings of the Royal Society of London. Series B: Biological Sciences 268, 745–753.

Nowak, M.A., 2006. Five rules for the evolution of cooperation. Science 314, 1560–1563.

Nowak, M.A., Roch, S., 2006. Upstream reciprocity and the evolution of gratitude. Proceedings of the Royal Society B: Biological Sciences 274, 605–610.

Nowak, M.A., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. Nature 393, 573.

Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness? Reputation dynamics in indirect reciprocity. Journal of Theoretical Biology 231, 107–120.

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

Okuno-Fujiwara, M., Postlewaite, A., 1995. Social norms and random matching games. Games and Economic Behavior 9, 79–109.

Panchanathan, K., Boyd, R., 2003. A tale of two defectors: The importance of standing for evolution of indirect reciprocity. Journal of Theoretical Biology 224, 115–126.

Prinz, J., 2007. The emotional construction of morals. Oxford University Press.

Resnick, P., Zeckhauser, R., Swanson, J., Lockwood, K., 2006. The value of reputation on ebay: A controlled experiment. Experimental Economics 9, 79–101.

Seinen, I., Schram, A., 2006. Social status and group norms: Indirect reciprocity in a repeated helping experiment. European Economic Review 50, 581–602.

Selten, R., 1967. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments, in: Sauermann, H. (Ed.), Beiträge zur Experimentellen Wirtschaftsforschung. Tübingen: J.C.B. Mohr. volume 1, pp. 136–168.

Sinclair, N., 2012. Metaethics, teleosemantics and the function of moral judgements. Biology & Philosophy 27, 639–662.

Sugden, R., 1986. The economics of rights, co-operation and welfare. Blackwell, Oxford, UK.

Swakman, V., Molleman, L., Ule, A., Egas, M., 2016. Reputation-based cooperation: empirial evidence for behavioral strategies. Evolution and Human Behavior 37, 230–235.

Takahashi, S., 2010. Community enforcement when players observe partners' past play. Journal of Economic Theory 145, 42–62.

Trivers, R.L., 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46, 35–57.

Weber, C.R., Federico, C.M., 2013. Moral foundations and heterogeneity in ideological preferences. Political Psychology 34, 107–126.

Yamamoto, H., Okada, I., Uchida, S., Sasaki, T., 2017. A norm knockout method on indirect reciprocity to reveal indispensable norms. Scientific Reports 7, 44146.

A Appendices

A.1 Proofs

Lemma.
Inserting the definition of $p_C(s)$ and $p_C(s,s')$ into the definition of profits given by equation (9) gives us:

$$\Pi_s(\tau;c,\eta)-\Pi_{s'}(\tau;c,\eta)=\sum_{\tilde{s}} x_{\tilde{s}}(\tau)\big(p_C(\tilde{s},s)-p_C(\tilde{s},s')\big)-\frac{c}{2}\big(p_C(s)-p_C(s')\big) \tag{A.1}$$

$$\begin{aligned}=\frac{1}{2}\sum_{\tilde{s}} x_{\tilde{s}}(\tau)\big(p_{\tilde{s}|CC}(p_{CC|s}-p_{CC|s'})+\ldots+p_{\tilde{s}|DD}(p_{DD|s}-p_{DD|s'})\big)\\ -\frac{c}{2}\big((p_{s|CC}-p_{s'|CC})p_{CC}+\ldots+(p_{s|DD}-p_{s'|DD})p_{DD}\big)\end{aligned} \tag{A.2}$$

Inserting equations (2)–(5),

$$\begin{aligned}=\frac{1}{2}\sum_{\tilde{s}} x_{\tilde{s}}(t)\big(p_{\tilde{s}|CC}(p_{s|CC}-p_{s'|CC})p_{CC}+\ldots+p_{\tilde{s}|DD}(p_{s|DD}-p_{s'|DD})p_{DD}\big)\\ -\frac{c}{2}\big((p_{s|CC}-p_{s'|CC})p_{CC}+\ldots+(p_{s|DD}-p_{s'|DD})p_{DD}\big)\end{aligned} \tag{A.3}$$

Inserting (1),

$$\begin{aligned}=\frac{b}{2}\sum_{\tilde{s}} x_{\tilde{s}}(\tau)\Big(p_{\tilde{s}|CC}\big((s_1-s_1')(1-\eta)+\ldots+(s_4-s_4')\frac{\eta}{3}\big)(1-\eta)p_{CC}+\ldots\\ +p_{\tilde{s}|DD}\big((s_4-s_4')(1-\eta)+\ldots+(s_1-s_1')\frac{\eta}{3}\big)(1-\eta)p_{DD}\Big)\\ -\frac{c}{2}\Big(\big((s_1-s_1')(1-\eta)+\ldots+(s_4-s_4')\frac{\eta}{3}\big)(1-\eta)p_{CC}+\ldots\\ +\big((s_4-s_4')(1-\eta)+\ldots+(s_1-s_1')\frac{\eta}{3}\big)p_{DD}\Big)\end{aligned} \tag{A.4}$$

Collecting the terms $(s_i - s_i')$,

$$
\begin{aligned}
=\frac{1}{2}\Bigg(& (s_1-s_1')(1-\eta)\Big((1-\eta)p_{CC}\sum_{\tilde{s}} x_{\tilde{s}}(\tau)p_{\tilde{s}|CC}+\ldots+\frac{\eta}{3}p_{DD}\sum_{\tilde{s}} x_{\tilde{s}}(\tau)p_{\tilde{s}|DD}\Big)+\ldots \\
& +(s_4-s_4')(1-\eta)\Big(p_{CC}\frac{\eta}{3}\sum_{\tilde{s}} x_{\tilde{s}}(\tau)p_{\tilde{s}|CC}+\ldots+(1-\eta)p_{DD}\sum_{\tilde{s}} x_{\tilde{s}}(\tau)p_{\tilde{s}|DD}\Big)\Bigg) \\
& -\frac{c}{2}\Bigg((s_1-s_1')(1-\eta)\Big((1-\eta)p_{CC}+\frac{\eta}{3}p_{CD}+\frac{\eta}{3}p_{DC}+\frac{\eta}{3}p_{DD}\Big)+\ldots \\
& +(s_4-s_4')(1-\eta)\Big(\frac{\eta}{3}p_{CC}+\frac{\eta}{3}p_{CD}+\frac{\eta}{3}p_{DC}+(1-\eta)p_{DD}\Big)\Bigg)
\end{aligned}
\tag{A.5}
$$

Applying the definition of the four basic payoff differences,

$$
=\sum_{i=1}^{4}(s_i-s_i')\Delta\Pi_i(\tau;c,\eta) \tag{A.6}
$$

□

Proposition.
The proof proceeds in two steps. In step one we show that for 26 of the 30 potentially stable equilibria presented in Table 1 there are non-equilibrium strategies yielding strictly higher payoffs contradicting asymptotic stability. The 30 candidates were derived with the help of Mathematica 10 (see supplementary). In step two we analyze each of the remaining equilibria separately.

Step One

1. $\{(x_s)|s\in S\}$
   For each of the 16 cases we analyze the profits of strategy s under the condition that $x_s=1$. It turns that population states with $x_s=1$ for $s\in\{(1,1,1,1),(1,0,1,1),(0,1,1,1),(1,0,1,0),(0,1,0,1),(0,0,1,1),(1,0,0,1),(0,0,1,0),(0,0,0,1)\}$ are vulnerable for invasions by all defectors, i.e., by strategy $(0,0,0,0)$.
   Let $\Delta\Pi_s=\big(\Pi_s(\tau,c,\eta)-\Pi_{(0,0,0,0)}(\tau,c,\eta)\big)\big|_{x_s=1}$. Payoff differences are given by
   $\Delta\Pi_{(1,1,1,1)}=-(1-\eta)$, $\Delta\Pi_{(1,0,1,1)}=-\frac{(3-\eta)(1-\eta)(12-31\eta+23\eta^2-4\eta^3+3c(7-4\eta))}{3c(9+12\eta-31\eta^2+23\eta^3-4\eta^4)}$,
   $\Delta\Pi_{(0,1,1,1)}=-\frac{(3-\eta)(1-\eta)((1-\eta)^2(9-15\eta+4\eta^2)+3c(6-7\eta+4\eta^2))}{3c(27-54\eta+55\eta^2-23\eta^3+4\eta^4)}$, $\Delta\Pi_{(1,0,1,0)}=-\frac{2(1-\eta)}{7-4\eta}$,
   $\Delta\Pi_{(0,1,0,1)}=-\frac{(1-\eta)(3-2\eta)}{6-7\eta+4\eta^2}$, $\Delta\Pi_{(0,0,1,1)}=-\frac{(1-\eta)(3-2\eta)(3+3c-7\eta+4\eta^2)}{3c(6-7\eta+4\eta^2)}$,
   $\Delta\Pi_{(1,0,0,1)}=-\frac{(1-\eta)(3-2\eta)(3+c(7-4\eta)-7\eta+4\eta^2)}{c(51-94\eta+68\eta^2-16\eta^3)}$, $\Delta\Pi_{(0,0,1,0)}=-\frac{(1-\eta)(9c+(1-\eta)^2\eta(3-4\eta)}{3c(24-22\eta+11\eta^2-4\eta^3)}$

$\Delta\Pi_{(0,0,0,1)} = -\frac{(1-\eta)^2(9+9c-24\eta+22\eta^2-11\eta^3+4\eta^4)}{c(27-45\eta+34\eta^2-11\eta^3+4\eta^4)}$. By inspection, $c \in (0,1)$ and $\eta \in (0, \frac{1}{5})$ imply that all these differences are strictly negative. Moreover, population states with $x_s = 1$ for $s \in \{(1,1,1,0),(1,1,0,1),(0,1,1,0)\}$ are also vulnerable to invasions.

$\Pi_{(1,1,1,0)} - \Pi_{(0,0,1,1)} = -\frac{(1-\eta)(3c(27-102\eta+143\eta^2-58\eta^3+8\eta^4)+2\eta(72-342\eta+589\eta^2-447\eta^3+144\eta^4-16\eta^5))}{9c(9+19\eta-23\eta^2+4\eta^3)}$,

$\Pi_{(1,1,0,1)} - \Pi_{(1,1,1,0)} = -\frac{(4-\eta)\eta^2(3-\eta-10\eta^2+8\eta^3)}{3(9+12\eta-31\eta^2+23\eta^3-4\eta^4)}$, $\Pi_{(0,1,0,0)} - \Pi_{(1,0,0,0)} = -\frac{(1-\eta)^2\eta^2(15-26\eta+8\eta^2)}{3(9+3\eta-10\eta^2+11\eta^3-4\eta^4)}$

$\Pi_{(0,1,1,0)} - \Pi_{(1,0,1,0)} = -\frac{2(1-\eta)^2\eta(3-4\eta)(3+3c-7\eta+4\eta^2)}{3c(3-2\eta)(11-10\eta+8\eta^2)}$. Again, all these differences are negative. Thus, the remaining candidates are the strategies $(1,1,0,0)$ and $(1,0,0,0)$. Regarding the former, note that $\Pi_{(1,1,0,0)} - \Pi_{(1,0,1,0)} = 0$ for population states with $x_{(1,1,0,0)} + x_{(1,0,1,0)} = 1$. Hence, the population state $x_{(1,1,0,0)} = 1$ cannot be stable as an equilibrium point. In the case of the latter, it turns out that either $(1,1,0,0)$ or $(0,0,0,0)$ earns strictly higher payoffs in a population state with $x_{(1,0,0,0)} = 1$. First, $\Pi_{(1,0,0,0)} - \Pi_{(1,1,0,0)} = -\frac{(1-\eta)\eta(2(1-\eta)^3(9-4\eta)(3-4\eta)-3c(36-\eta(53-34\eta+8\eta^2)))}{9c(18-\eta(2+\eta(11-4\eta)))}$. Second, $\Pi_{(1,0,0,0)} - \Pi_{(0,0,0,0)} = -\frac{(1-\eta)\eta(21-12\eta-((1-\eta)^2(3-4\eta))/c)}{54-3\eta(2+\eta(11-4\eta))}$. In sum, only the population state $x_{(0,0,0,0)} = 1$ may constitute a stable equilibrium point.

2. $(x_{(1,1,0,1)}, x_{(0,1,0,1)})$

   It turns out that under the assumption of equal profits for strategies $(1,1,0,1)$ and $(0,1,0,1)$ a state characterized by $x_{(1,1,0,1)} + x_{(0,1,0,1)} = 1$ is vulnerable to an invasion by strategy $(1,1,1,0)$. Under the restriction to $\eta \in (0, \frac{1}{5})$, the condition $x_{(1,1,0,1)} \in (0,1)$ reduces to $c \in \big(0, \frac{(3-\eta)(1-\eta)^2(3-4\eta)(9-(3-\eta)\eta(7-4\eta))}{3(27-\eta(99-\eta(174-\eta(143-58\eta+8\eta^2))))}\big)$. In this range of cost-benefit ratios the payoff difference $\Pi_{(1,1,0,1)} - \Pi_{(1,1,1,0)}$ is strictly negative.

3. $(x_{(1,0,0,0)}, x_{(0,0,0,0)})$

   It turns out that under the assumption of equal profits for strategies $(1,0,0,0)$ and $(0,0,0,0)$ a state characterized by $x_{(1,0,0,0)} + x_{(0,0,0,0)} = 1$ is vulnerable to an invasion by strategy $(1,1,0,0)$. Under the restriction to $\eta \in (0, \frac{1}{5})$, the condition $x_{(1,0,0,0)} \in (0,1)$ reduces to $c \in \big(0, \frac{(1-\eta)^2(3-4\eta)}{3(7-4\eta)}\big)$. In this range of cost-benefit ratios the payoff difference $\Pi_{(1,0,0,0)} - \Pi_{(1,1,0,0)}$ is strictly negative.

4. $(x_{(1,1,0,1)}, x_{(1,0,0,1)})$

   It turns out that under the assumption of equal profits for strategies $(1,1,0,1)$ and $(1,0,0,1)$ a state characterized by $x_{(1,1,0,1)} + x_{(1,0,0,1)} = 1$ is vulnerable to an invasion by strategy $(1,1,1,0)$. Under the restriction to $\eta \in (0, \frac{1}{5})$, the condition $x_{(1,1,0,1)} \in (0,1)$ reduces to $c \in \big(\frac{(1-\eta)(3-4\eta)(45-2\eta(81-2\eta(46-\eta(23-4\eta))))}{9(15-\eta(31-\eta(23-4\eta)))}, \frac{(3-\eta)(1-\eta)^3(15-4\eta)(3-4\eta)}{9(15-\eta(31-\eta(23-4\eta)))}\big)$. In this range of cost-benefit ratios the payoff difference $\Pi_{(1,1,0,1)} - \Pi_{(1,1,1,0)}$ is strictly negative.

5. $(x_{(0,1,0,0)}, x_{(0,0,0,0)})$
   It turns out that under the assumption of equal profits for strategies $(0,1,0,0)$ and $(0,0,0,0)$ a state characterized by $x_{(0,1,0,0)} + x_{(0,0,0,0)}=1$ is vulnerable to an invasion by strategy $(1,1,0,1)$. Under the restriction to $\eta\in(0,\frac{1}{5})$, the condition $x_{(0,1,0,0)}\in(0,1)$ reduces to $c\in\big(0, \frac{(1-\eta)(3-4\eta)(3-\eta(1-\eta))}{3(6-\eta(7-4\eta))}\big)$. In this range of cost-benefit ratios the payoff difference $\Pi_{(0,1,0,0)} - \Pi_{(1,1,0,1)}$ is strictly negative.

6. $(x_{(1,1,1,1)}, x_{(1,1,0,1)})$
   It turns out that there are parameter constellations, for instance, $(c=\frac{1}{8}, \eta=\frac{3}{32})$ such that the equilibrium strategies $(1,1,1,1)$ and $(1,1,0,1)$ earn the same and strictly highest payoff in a state characterized by $x_{(1,1,1,1)} + x_{(1,1,0,1)}=1$. Thus, this equilibrium is a candidate to be considered in step two.

7. $(x_{(1,1,1,0)}, x_{(1,1,0,0)})$
   It turns out that under the assumption of equal profits for strategies $(1,1,1,0)$ and $(1,1,0,0)$ a state characterized by $x_{(1,1,1,0)} + x_{(1,1,0,0)}=1$ is vulnerable to an invasion by strategy $(0,0,0,0)$. Under the restriction to $\eta\in(0,\frac{1}{5})$, the condition $x_{(1,1,1,0)}\in(0,1)$ reduces to $c\in\big(\frac{(1-\eta)(3-4\eta)(4-\eta)(9-2\eta(7-4\eta))}{3(45-2\eta(37-\eta(23-4\eta)))}, \frac{(1-\eta)(3-4\eta)}{3}\big)$. In this range of cost-benefit ratios the payoff difference $\Pi_{(1,1,1,0)} - \Pi_{(0,0,0,0)}$ is strictly negative.

8. $(x_{1,0,1,0}, x_{1,0,0,0})$
   It turns out that under the assumption of equal profits for strategies $(1,0,1,0)$ and $(1,0,0,0)$ a state characterized by $x_{(1,0,1,0)} + x_{(1,0,0,0)}=1$ is vulnerable to an invasion by strategy $(1,1,0,1)$. Under the restriction to $\eta\in(0,\frac{1}{5})$, the condition $x_{(1,1,1,0)}\in(0,1)$ reduces to $c\in\big(\frac{(1-\eta)^2\eta(3-4\eta)(7-4\eta)}{3(36-\eta(53-34\eta+8\eta^2)}, \frac{576}{3-4\eta} - 21 - \frac{27}{(1-\eta)^2} - \frac{144}{1-\eta} + 26\eta - 8\eta^2 - \sqrt{\theta}\big)$, where $\theta=\frac{3\eta^2(7-4\eta)^3(10-\eta(11-4\eta))(6-\eta(7-4\eta))}{(3-4\eta)^2(1-\eta)^4}$. In this range of cost-benefit ratios the payoff difference $\Pi_{(1,0,1,0)} - \Pi_{(1,1,0,1)}$ is strictly negative.

9. $(x_{0,1,1,0}, x_{0,1,0,0})$
   It turns out that under the assumption of equal profits for strategies $(0,1,1,0)$ and $(0,1,0,0)$ a state characterized by $x_{(0,1,1,0)} + x_{(0,1,0,0)}=1$ is vulnerable to an invasion by strategy $(0,0,0,0)$. Under the restriction to $\eta\in(0,\frac{1}{5})$, the condition $x_{(1,1,1,0)}\in(0,1)$ reduces to $c\in\big(\frac{(1-\eta)(3-4\eta)(3-4\eta(1-\eta))}{9}, \frac{(1-\eta)(3-4\eta)(3-\eta(1-\eta))}{9}\big)$. In this range of cost-benefit ratios the payoff difference $\Pi_{(0,1,1,0)} - \Pi_{(0,0,0,0)}$ is strictly negative.

10. $(x_{1,1,1,1}, x_{1,1,1,0})$
    It turns out that under the assumption of equal profits for strategies $(1,1,1,1)$ and $(1,1,1,0)$ a state characterized by $x_{(1,1,1,1)} + x_{(1,1,1,0)}=1$ is vulnerable to an

invasion by strategy $(0,0,0,0)$. Under the restriction to $\eta\in(0,\frac{1}{5})$, the condition $x_{(1,1,1,1)}\in(0,1)$ reduces to $c\in\big(0,\frac{4(1-\eta)(3-4\eta)(4-\eta)}{57-12\eta}\big)$. In this range of cost-benefit ratios the payoff difference $\Pi_{(1,1,1,1)}-\Pi_{(0,0,0,0)}$ is strictly negative.

11. $(x_{(1,1,1,0)},x_{(1,0,1,0)},x_{(0,1,1,0)},x_{(0,0,1,0)})$
It turns out that under the assumption of equal profits for strategies $x_{(1,1,1,0)}$, $x_{(1,0,1,0)}$, $x_{(0,1,1,0)}$, and $x_{(0,0,1,0)}$ a state characterized by $x_{(1,1,1,0)}+x_{(1,0,1,0)}+x_{(0,1,1,0)}+x_{(0,0,1,0)}=1$ is vulnerable to an invasion by strategy $(1,1,0,1)$. See supplementary for details.

12. $(x_{(1,1,0,1)},x_{(0,1,0,1)},x_{(1,0,0,1)},x_{(0,0,0,1)})$
It turns out that under the assumption of equal profits for strategies $x_{(1,1,0,1)}$, $x_{(0,1,0,1)}$, $x_{(1,0,0,1)}$, and $x_{(0,0,0,1)}$ a state characterized by $x_{(1,1,0,1)}+x_{(0,1,0,1)}+x_{(1,0,0,1)}+x_{(0,0,0,1)}=1$ is vulnerable to an invasion by strategy $(0,0,0,0)$. See supplementary for details.

13. $(x_{(1,1,1,1)},x_{(1,1,1,0)},x_{(1,1,0,1)},x_{(1,1,0,0)})$
It turns out that there exist parameter constellations, for instance $(c=\frac{1}{3},\eta=\frac{1}{8})$ such that the equilibrium strategies earn the same and strictly highest payoffs in a state characterized by $x_{(1,1,1,1)}+x_{(1,1,1,0)}+x_{(1,1,0,1)}+x_{(1,1,0,0)}=1$. Thus, this equilibrium set is a candidate to be considered in step two.

14. $(x_{(0,1,1,1)},x_{(0,1,1,0)},x_{(0,1,0,1)},x_{(0,1,0,0)})$
It turns out that under the assumption of equal profits for strategies $x_{(0,1,1,1)}$, $x_{(0,1,1,0)}$, $x_{(0,1,0,1)}$, and $x_{(0,1,0,0)}$ a state characterized by $x_{(0,1,1,1)}+x_{(0,1,1,0)}+x_{(0,1,0,1)}+x_{(0,1,0,0)}=1$ is vulnerable to an invasion by strategy $(0,0,0,0)$. See supplementary for details.

15. $(x_{(1,1,1,1)},...,x_{(0,0,0,0)})$
By definition, in this equilibrium all 16 earn the same payoff. Consequently, there cannot exist a strategy that yields higher payoffs than the equilibrium strategies. However, there exists a subset of this equilibrium set which is unstable such that the neutral drift will eventually shift the population state into that region. The set of population states which satisfy $\Delta\Pi_i(\tau;c,\eta)=0,\forall i$ is given by $\big\{x\in\Delta_{15}\big|\sum_{s\in S}x_s=1,\ x_{(0,1,0,0)}=\gamma+\sum_{s\in S_2}x_s-\sum_{s\in S_3}x_s,\ x_{(0,0,1,0)}=-\gamma+\sum_{s\in S_4}x_s-\sum_{s\in S_5}x_s\big\}$, where $\gamma=\frac{3c}{(1-\eta)(3-4\eta)}$, $S_2=\{x_{(1,0,1,1)},x_{(0,0,1,1)},x_{(1,0,0,1)},x_{(0,0,0,1)}\}$, $S_3=\{x_{(1,1,1,0)},x_{(1,1,0,0)},x_{(0,1,1,0)}\}$, $S_4=\{x_{(1,1,0,1)},x_{(1,1,0,0)},x_{(1,0,0,1)},x_{(1,0,0,0)}\}$, $S_5=\{x_{(0,1,1,1)},x_{(0,0,1,1)},x_{(0,1,1,0)}\}$. Consider,

for instance, the solution $x_{(1,1,0,0)}=\frac{3c}{3-7\eta+4\eta^2}, x_{(0,0,0,0)}=1-\frac{3c}{3-7\eta+4\eta^2}, x_s=0, \forall s\in S\setminus\{(1,1,0,0),(0,0,0,0)\}$ which satisfies the conditions for existence of this equilibrium, i.e., it solves $\Delta\Pi_i(\tau;c,\eta)=0, \forall i$. It turns out that with the exception of one strictly positive eigenvalue all other eigenvalues are zero. This eigenvalue is given by $\frac{2c\eta(3(1-c)-7\eta+4\eta^2)}{2(1-\eta)(3-4\eta)}$.

Taken together there are three candidates for stable (sets of) population states. The stability of those candidates is analyzed in detail below.

Step Two

1. $x_{(0,0,0,0)}$
   This equilibrium is stable for all pair of parameters, i.e., for $\frac{c}{b}\in(0,1)$ and $\eta\in(0,\frac{1}{5})$. This is because all other strategies earn strictly negative payoffs while $(0,0,0,0)$ earns a payoff of zero. This results from the assumptions that there is no unintended help but only unintended defection and that there are observational errors. Thus, any strategy which prescribes cooperation for some label will eventually induce cooperation because either this label is present in the population or some other is misperceived as this particular label. As a consequence, participants playing a strategy different from $(0,0,0,0)$ will bear the cost of cooperation with some probability but will never receive help in a state with $x_{(0,0,0,0)}=1$.

2. $(x_{(1,1,1,1)}, x_{(1,1,0,1)})$
   Note first that this equilibrium point is internally stable in the sense that an increase in the share of $x_{(1,1,1,1)}$ decreases the payoffs of $(1,1,1,1)$ relative to $(1,1,0,1)$ (see supplementary). Second, as a necessary condition for stability we must have that $\Delta\Pi_i(t;c,\eta)>0, i\neq 3$. This guarantees that a deviation from cooperation for labels CC, CD, or DD is not profitable. Taken together with the requirement that $x_{(1,1,1,1)}\in(0,1)$ this reduces to $\frac{c}{b}\in\big(0,\frac{3-10\eta+8\eta^2}{15-22\eta+8\eta^2}\big)$ and $\eta\in(0,\frac{1}{5})$. To establish sufficiency we study the eigenvalues of the linearized system at this equilibrium point. Analytic solutions are presented in the supplementary. It turns out that the conditions $\frac{c}{b}\in\big(0,\frac{3-10\eta+8\eta^2}{15-22\eta+8\eta^2}\big)$ and $\eta\in(0,\frac{1}{5})$ imply that all eigenvalues are strictly negative, which establishes the sufficiency of these conditions.

3. $(x_{(1,1,1,1)}, x_{(1,1,1,0)}, x_{(1,1,0,1)}, x_{(1,1,0,0)})$
   The condition of being well-defined, i.e., $(x_{(1,1,1,1)}, x_{(1,1,1,0)}, x_{(1,1,0,1)}, x_{(1,1,0,0)})\in\Delta_3$, reduces with respect $\frac{c}{b}$ and η to $\frac{c}{b}\in\big[\frac{3-10\eta+8\eta^2}{15-22\eta+8\eta^2}, \frac{12-52\eta+72\eta^2-32\eta^3}{15-22\eta+8\eta^2}\big]$ and $\eta\in(0,\frac{1}{5})$. As

a necessary condition for stability we must have that $\Delta\Pi_i(t;c,\eta)>0, i=1,2$. This guarantees that a deviation from cooperation for labels CC, or CD is not profitable. Taken together with the condition for existence we obtain $\frac{c}{b}\in\left[\frac{3-10\eta+8\eta^2}{15-22\eta+8\eta^2}, \frac{3-10\eta+8\eta^2}{6-4\eta}\right)$ and $\eta\in(0,\frac{1}{5})$. Note that at the lower bound of $\frac{c}{b}$ the equilibria (2) and (3) coincide. That is, the equilibrium set reduces to an equilibrium point with $x_{(1,1,1,1)}+x_{(1,1,0,1)}=1$.

To establish sufficiency we study the eigenvalues of the linearized system at a point in this equilibrium. It turns out that there is a eigenvalue of zero with multiplicity one, three eigenvalues with multiplicity four, and two non-zero eigenvalues with multiplicity one. The eigenvalue of zero with multiplicity one corresponds to the dimension of the equilibrium set and reflects the vanishing payoff differences for the equilibrium strategies. In other words, the corresponding eigenvector $(1,-1,-1,0,0,1,0,...,0)'$ spans this equilibrium set. Note that the equilibrium set connects two facets of 3-simplex. Particularly, at the population state with the minimal share of unconditional cooperators $\left(x_{(1,1,1,1)}=\frac{\eta(1+2\eta)}{3-5\eta+2\eta^2}\right)$ we have $x_{(1,1,0,0)}=0$, at the population state that maximizes the share of unconditional cooperators $\left(x_{(1,1,1,1)}=1+\frac{4}{3-4\eta}-\frac{8}{3-2\eta}+\frac{3c(5-4\eta)}{(3-4\eta)^2(1-\eta)}\right)$ the share $x_{(1,1,1,0)}$ vanishes. Thus, the eigenvalue of zero does not cause any issue of instability at the boundaries of this equilibrium set.

The three eigenvalues with multiplicity four are: $e_{2-5}=\frac{2(1-\eta)^2(6-4\eta-\frac{3-2\eta(5-4\eta)}{c})}{(3-2\eta)(3+6\eta-8\eta^2)}$, $e_{6-9}=\frac{2(1-\eta)^2\eta(6-4\eta-\frac{3-2\eta(5-4\eta)}{c})(5-4\eta)}{3(3-2\eta)(3+6\eta-8\eta^2)}$, and $e_{10-13}=\frac{2(1-\eta)^2(6-4\eta-\frac{3-2\eta(5-4\eta)}{c})(3-5\eta-4\eta^2)}{3(3-2\eta)(3+6\eta-8\eta^2)}$. Note that eigenvalues 2–13 are linear in $\frac{1}{c}$ and do not depend on the chosen element of the equilibrium set. It turns out that they all share the same root. That is, eigenvalues 2–13 are strictly negative if and only if $\frac{c}{b}<\frac{3-10\eta+8\eta^2}{6-4\eta}$. The eigenvalues e_{14} and e_{15} are more complicated expressions, in particular they do depend on the point of the equilibrium set under consideration (see supplementary). However, under the necessary condition $\Delta\Pi_i(t;c,\eta)>0, i=1,2$ for any well-defined element of this equilibrium set eigenvalues e_{14} and e_{15} are also strictly negative. Thus, the necessary condition is also sufficient for stability.

□

A.2 Dynamics of labels

In general, the transition matirx $(w)_{ij}$ is given by:

$$\begin{pmatrix} \sum_{s\in S} x_s p_{s|CC} & \sum_{s\in S} x_s p_{s|CD} & 0 & 0 \\ 0 & 0 & \sum_{s\in S} x_s p_{s|DC} & \sum_{s\in S} x_s p_{s|DD} \\ 1-\sum_{s\in S} x_s p_{s|CC} & 1-\sum_{s\in S} x_s p_{s|CD} & 0 & 0 \\ 0 & 0 & 1-\sum_{s\in S} x_s p_{s|DC} & 1-\sum_{s\in S} x_s p_{s|DD} \end{pmatrix} \tag{A.7}$$

Making use of the identity $p_{CC}(t)+p_{CD}(t)+p_{DC}(t)+p_{DD}(t)=1,\ \forall t$, the recursive system (7) simplifies to $(p_{CC}(t),p_{CD}(t),p_{DC}(t))^T=(\tilde{w})_{ij}\cdot(p_{CC}(t-1),p_{CD}(t-1),p_{DC}(t-1))^T$:

$$(\tilde{w})_{ij}=\begin{pmatrix} \tilde{w}_{11} & \tilde{w}_{12} & 0 \\ \tilde{w}_{21} & \tilde{w}_{22} & \tilde{w}_{23} \\ \tilde{w}_{31} & \tilde{w}_{32} & 0 \end{pmatrix} \tag{A.8}$$

$\tilde{w}_{ij}$ are functions of the error η and the population state. The recursive system above is stable if and only if the absolute value of all eigenvalues is below unity. Without any restrictions the eigenvalues are analytically not tractable. Note, however, that for $\eta=0$ $(\tilde{w})_{ij}$ simplifies to

$$(\tilde{w})_{ij}=\begin{pmatrix} \sum_{i\in S_{11}} x_i & \sum_{i\in S_{12}} x_i & 0 \\ \sum_{i\in S_{21}} x_i & \sum_{i\in S_{21}} x_i & \sum_{i\in S_{23a}} x_i - \sum_{i\in S_{23b}} x_i \\ 1-\sum_{i\in S_{11}} x_i & 1-\sum_{i\in S_{12}} x_i & 0 \end{pmatrix}, \tag{A.9}$$

where
$S_{11}=\{(1,1,1,1),(1,1,1,0),(1,1,0,1),(1,0,1,1),(1,1,0,0),(1,0,1,0),(1,0,0,1),(1,0,0,0)\}$,
$S_{12}=\{(1,1,1,1),(1,1,1,0),(1,1,0,1),(0,1,1,1),(1,1,0,0),(0,1,1,0),(0,1,0,1),(0,1,0,0)\}$,
$S_{21}=\{(1,1,1,1),(1,1,0,1),(1,0,1,1),(0,1,1,1),(0,1,0,1),(0,0,1,1),(1,0,0,1),(0,0,0,1)\}$,
$S_{23a}=\{(1,1,1,0),(1,0,1,0),(0,1,1,0),(0,0,1,0)\}$,
and $S_{23b}=\{(1,1,0,1),(0,1,0,1),(1,0,0,1),(0,0,0,1)\}$.

Note further that both cooperative equilibria satisfy $x_{(1,1,1,1)}+x_{(1,1,1,0)}+x_{(1,1,0,1)}+x_{(1,1,0,0)}=1$. Under this condition eigenvalues of $(\tilde{w})_{ij}$ are given by $\lambda_1=\lambda_2=0$ and $\lambda_3=1-x_{(1,1,1,1)}-x_{(1,1,0,1)}$. Hence, due to the continuity of the eigenvalues in η, the fixed point of the label distribution given by (8) is a global attractor for $\eta\approx 0$ and for an equilibrium population state. This is because $\eta>0$ implies $x_{(1,1,1,1)}>0$ for both cooperative equilibria (see Proposition). We extend this results numerically to $\eta\leq 1/10$. For the details, we refer

the reader to the supplementary material.

For arbitrary population states we have to rely on simulations which can be found in the supplementary material. The simulation results show that for any population state and any initial distribution of labels, the actual distribution is very close to the fixed point given by (8) after less than 5 periods.

A.3 Distribution of labels over time, by treatment

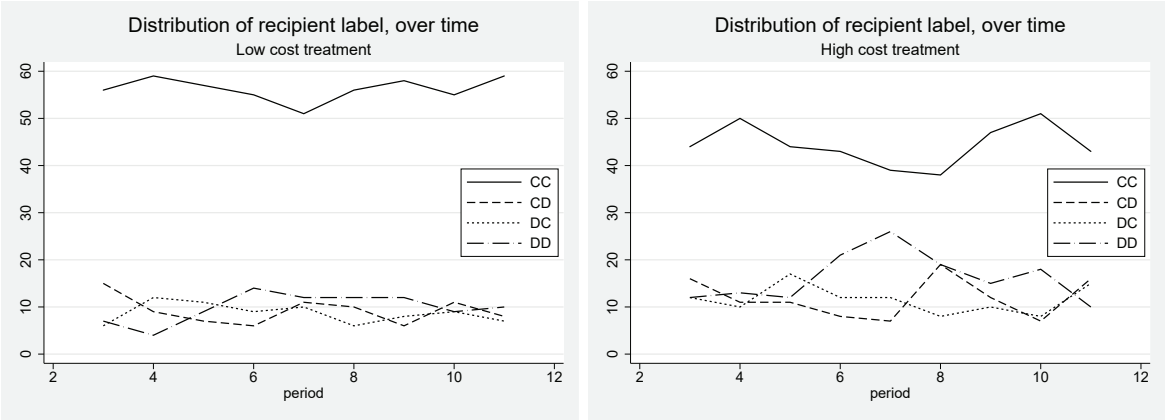

Figure 4: Distribution of labels over time, by treatment.

A.4 Errors of implementation

	Period 4 vs. period 5	Period 10 vs. period 11	Total
Low cost treatment	4.8% (63)	9.0% (67)	6.9% (130)
High cost treatment	5.5% (55)	10.7% (56)	8.1% (111)
Total	5.1% (118)	9.8% (123)	7.5% (241)

Table 6: Error rates, by treatment, in percentage (N in parenthesis).
Error rates are the frequency with which a person who stated C for a given label in period 5 or 11 did D for that label in period 4 or 10.

A.5 Choice of strategies

We run multinomial logit regressions with six categories, one for each equilibrium strategy, one for $(0, 0, 0, 0)$, and one for all other strategies. The reference category is $(1, 1, 1, 1)$. We select variables by the Akaike Information Criterion in a stepwise algorithm both

adding and removing terms, starting with the full model containing all information from the questionnaire. The remaining variables are *consistency*, a dummy which is one if a participant chose the same strategy in periods five and eleven; *similar*, a dummy which is one if a participant reported feeling similar to the other participants; *CRT score*, the number of correct answers in the CRT questions, *male*, a dummy for a participant's gender, and *age*, the participant's age in years. We report regressions without 6 observation corresponding to participants who did not indicate their gender, but results are robust to including those observations. The results are presented in Table 7.

	(0,0,0,0)	(1,1,0,0)	(1,1,0,1)	(1,1,1,0)	other
(Intercept)	0.05(2.21)	0.22(1.40)	−0.29(2.83)	−1.50(1.55)	−1.83(1.54)
consistency	0.61(0.77)	0.43(0.53)	−0.32(0.73)	−1.08(0.59)	−1.22(0.60)*
similar	−1.55(0.62)*	0.55(0.49)	0.76(0.76)	0.27(0.63)	−0.46(0.60)
CRT score	−0.41(0.28)	−0.22(0.22)	1.15(0.39)**	−0.08(0.30)	−0.20(0.30)
age	−0.07(0.09)	0.01(0.05)	−0.14(0.11)	0.08(0.05)	0.10(0.05)*
male	0.94(0.64)	−0.17(0.45)	−1.06(0.68)	−0.90(0.62)	−1.53(0.68)*
AIC	537.19				
BIC	629.82				
Log Likelihood	−238.59				
Deviance	477.19				
Num. obs.	162				

$^{**}p<0.01$; $^{*}p<0.05$

Table 7: Multinomial logit regression. Standard errors in parentheses.

A.6 Instructions

General instructions

1. Welcome and thank you for your participation! At the beginning of the experiment you pulled out a cabin number and an envelope from a basket. PLEASE DO NOT OPEN THE ENVELOPE! We will let you know when you can open the envelope. This will happen only when the experiment is over and you have done ALL the necessary tasks. If you violate this rule and open your envelope beforehand, we will unfortunately be forced to exclude you from the experiment.

2. You can earn an amount of money in this experiment that depends on your choices or those of another person. It is therefore very important that you read all the instructions thoroughly and completely. Please switch off your mobile phones now!

Communication with other participants is not allowed. If you have a question, please raise your hand. We will then come to you and answer your question.

3. In this experiment, as with all experiments in the Göttingen Laboratory of Behavioral Economics (GLOBE), the participants shall not be deceived. If you are in any doubt that the instructions correspond to the truth, then please contact lab@uni-goettingen.de or a member of the Department of Microeconomics (Prof. Keser). All data from this study is stored anonymously and kept strictly confidential.

4. Your earnings are calculated in ECU (Experimental Currency Units). 1 ECU equals 0.10 €. At the end of today's session, your total earnings will be converted into euros and paid out confidentially and in cash. In addition, you will receive a participation fee of 4 € (40 ECU).

5. You and other participants are part of an experiment. This experiment goes on for several periods. In each period you will be assigned to interact with another person in today's experiment. You will not meet that person a second time. You will never know who the other people you meet are. The other people will not know your identity.

6. In the experiment you are either a "decision-maker" or a "recipient". The decision-maker can choose between "keep" and "give." If the decision-maker chooses "keep", the decision-maker receives 85 ECU and the recipient receives 40 ECU. If the decision-maker chooses "give," the decision-maker receives 75 ECU and the recipient receives 65 ECU.

7. A paper in your envelope states whether you are the decider or the recipient. If you are a decision-maker, the paper says "You are the decider." The paper also says which period will be relevant for payment. Every period can be relevant, but only ONE period will actually be relevant for your payoff. Your decision in the payoff-relevant period determines your payoff and the payoff of the person you met in that period. If you are a recipient, the envelope says "You are a recipient." Then the decision of another person who was the decision-maker in the payoff-relevant period will determine your payoff.

8. Every participant in this lab can be a decision-maker. Therefore, you should always make your decisions as if you were a decision-maker and as if you were deciding both your payoff and that of another person. You can not change your decision at

the end of the experiment. We will only implement the decision you made during the experiment.

9. You are now asked to answer six control questions to check your understanding of the instructions. The experiment will not begin until all participants have correctly answered the control questions.

Control questions

1. Which period will be payoff-relevant? (The first period; Period 9; None; I do not know that in advance. It's in my envelope.)

2. Will your decisions be payoff-relevant or will decisions made by another person determine your payoff? (My decisions will be payoff-relevant; The decisions of another person will determine my payoff; I do not know that in advance. It's in my envelope.)

3. How many times will you meet the same person in this experiment? (Once; Every period; I do not know that in advance.)

4. Will you know which person in the lab you meet in a given period? (Yes; No)

5. Will your identity be revealed to other people in the lab? (Yes; No)

6. Can the experimenters link your decisions with your name? (Yes; No)

Game instructions

1. The experiment begins now.

2. *Periods 1 to 4 and 6 to 10*:
   In period X, you meet someone new. You have two alternatives to choose from: "keep" or "give." If you choose "keep," you will receive 85 ECU and the other person will receive 40 ECU. If you choose "give", you will receive 75 ECU and the other person will receive 65 ECU.
   [Period 2: The other person has chosen (give/keep) in the last period.
   Period 3 to 4 and 6 to 10: The other person has chosen (give/keep) in the last period, when faced with someone whom they knew had chosen (give/keep).]
   Which alternative do you choose? (give/keep)
   Info: In the next stage, you will be the recipient and another person will be the

decision-maker. This person will know which choice you have made.
[Period 2 to 4 and 6 to 10: Info: In the next stage, you will be the recipient and another person's decision-maker. This person will know which choice you have made. In addition, this person will learn what the person you have met this turn had chosen before.]

3. Feedback: You have chosen (give/keep). The decision maker who meets you now learns that you have chosen (give/keep).
[Period 2 to 4 and 6 to 10: You have chosen (give/keep). The decision maker who meets you now learns that you have chosen (give/keep) when faced with someone who had chosen (give/keep).]

4. *Control questions before period 3:*
Suppose that one person has chosen (give/keep) when faced with someone whom they knew had chosen (give/keep). Which of the following statements are true?

   (a) Question 1: That person had chosen (give/keep) last period.

   (b) Question 2: That person had chose (give/keep) two periods ago.

   (c) Question 3: The person who was meeting that person had chosen (give /keep) two periods ago.

5. *Periods 5 and 11*:
In period X, we ask you about your decision plan for this period. We ask for your decision when faced with someone you meet during this period. You will see various possible information about the other person's decision. For each case you have to make a decision. Only at the end of the experiment, and if this period X is relevant, will we give you information about the actual decision of the other person. Your decision for the relevant case will then be executed and you will receive the appropriate payoff. Please make sure that your decisions are in line with what you want to see done at the end of the experiment!
You have two alternatives to choose from: "keep" or "give." If you choose "keep," you will receive 85 ECU and the other person will receive 40 ECU. If you choose "give," you will receive 75 ECU and the other person will receive 65 ECU.
(The order of the following questions was randomized)

   (a) Suppose that person chose to give in the last period, when faced with someone they knew had chosen to give before. Which alternative would you choose in this case?

(b) Suppose that person chose to give in the last period, when faced with someone they knew had chosen to keep before. Which alternative would you choose in this case?

(c) Suppose that person chose to keep in the last period, when faced with someone they knew had chosen to give before. Which alternative would you choose in this case?

(d) Suppose that person chose to keep in the last period, when faced with someone they knew had chosen to keep before. Which alternative would you choose in this case?

Payoffs

1. The game is now over. An experimenter will go through the lab to make sure none of the envelopes have been opened. Please wait until we let you know that you can open your envelope.

2. In your envelope is a paper telling you whether you are a decision-maker and which period is relevant to your payoff. Please enter "Yes," if your envelope states that you are the decider. Please also enter the period that appears in your envelope. The experimenter will come to you and will confirm your input so that you get to the last part of the experiment.

3. **Payoffs for the decision-maker:**
   You are a decision-maker. Your decision in period X determines your payoff and the other person's payoff.
   Period 1: In period 1 you chose (give/keep). **Period 2:** In period 2 you chose (give/keep). You knew that the other person had chosen (give/keep) in period 1. **Periods 3 to 11:** In period X you chose (give/keep). You knew that the other person had chosen (give/keep) in period X-1, when faced with someone you knew had chosen (give /keep) in the previous period.
   Therefore, you will receive XXX ECU and the other person will receive XXX ECU. Converted into euros and with a participation fee of 4€, you will get XXX euros. The other person will get XXX Euro.

4. **Payoffs for the recipient:**
   You are a recipient. The decision of another person in period 1 determines your payoff and the payoff of that person.

Period 1: In period 1 the other person has chosen (give/keep). **Period 2:** In period 2 the other person has chosen (give/keep). This person knew that you had chosen (give/keep) in period 1. **Periods 3 to 11:** In period X the other person has chosen (give /keep). This person knew that you had chosen (give/keep) in the period X-1, when faced with someone you knew had chosen (give/keep) in the previous period.
Therefore, you will receive XXX ECU and the other person will receive XXX ECU. Converted into euros and with a participation fee of 4€, you will get XXX euros. The other person will get XXX Euro.

The payoff stage is now over. We still ask you to answer a number of survey questions, after which you will be able to collect your remuneration.

Questionnaire

1. What do you think the experiment was about?
2. Please describe by which criteria you made decisions.
3. Do you think that the information provided to you was sufficient to make your decisions? If not, what else would you have liked to know before you made your decisions?
4. Was it hard for you to understand what you had to do in this experiment?
5. Did you have any difficulties during the experiment? If yes, please describe your difficulties.
6. In how many experiments did you already participate in (approximately)? 1. I have never participated in an experiment before. 2. 1-5 3. More than 5.
7. What kind of interactions did the experiment most likely remind you of? Interactions with 1. family members. 2. friends. 3. classmates/work colleagues. 4. strangers.
8. Do you know one or more of the other people who participated in today's experiment?
9. Do you think that you are similar to the other participants in today's experiment?

10. Anna's father has 5 daughters: Lala, Lele, Lili, Lolo, and .... What is the name of the fifth daughter?

11. On a boat hangs a ladder with five rungs. The distance from one rung to the next is 20cm. The lowest rung touches the water surface. The tide raises the water level by 20cm per hour. How long does it take for the water level to reach the topmost rung? 1. 5 hours. 2. 4 hours. 3. It never reaches it.

12. If it takes 20 minutes to boil a goose egg, how many minutes will it take to boil 3 geese eggs? 1. One hour. 2. 20 minutes.

13. Linda is 31 years old, single, very intelligent, and openly speaks her mind. She studied philosophy. During her studies, she dealt extensively with issues of equal rights and social justice and also took part in anti-nuclear demonstrations. Which statement is most likely to apply to Linda? 1. Linda is a bank clerk. 2. Linda is a bank clerk and is active in a feminist movement.

Finally, we would like to have some information about you.

1. How old are you? Please type in.

2. What is your gender?

3. What is your living situation? 1. I do not live alone. 2. I live alone.

4. Do you live in Göttingen?

5. Do you have German citizenship?

6. Are you worried about covering your living expenses over the next six months?

7. Do you think you are financially better off than others in your age group?

8. Are you currently enrolled for a degree?

9. Did you study economics in the past or are you now studying economics?

We thank you for your participation! You will now be called individually to receive your payment. Please come to the experimenter in the entrance area when your cabin number is called up. Please make sure you take all your belongings with you so you do not have to go back to the cabin.