Evolution and Human Behavior 34 (2013) 139–145

Contents lists available at SciVerse ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

Charitable giving as a signal of trustworthiness: Disentangling the signaling benefits of altruistic acts

Sebastian Fehrler [a,b], Wojtek Przepiorka [c,d,*]

[a] Department of Political Science, University of Zurich and Center for Comparative and International Studies (CIS), Switzerland
[b] Institute for the Study of Labor (IZA), Bonn, Germany
[c] Department of Sociology, University of Oxford, United Kingdom
[d] Chair of Sociology, ETH Zurich, Switzerland

ARTICLE INFO

Article history:
Initial receipt 21 July 2011
Final revision received 12 November 2012

Keywords:
Altruism
Evolution of cooperation
Costly signaling
Social preferences
Trust
Trustworthiness

ABSTRACT

It has been shown that psychological predispositions to benefit others can motivate human cooperation and the evolution of such social preferences can be explained with kin or multi-level selection models. It has also been shown that cooperation can evolve as a costly signal of an unobservable quality that makes a person more attractive with regard to other types of social interactions. Here we show that if a proportion of individuals with social preferences is maintained in the population through kin or multi-level selection, cooperative acts that are truly altruistic can be a costly signal of social preferences and make altruistic individuals more trustworthy interaction partners in social exchange. In a computerized laboratory experiment, we test whether altruistic behavior in the form of charitable giving is indeed correlated with trustworthiness and whether a charitable donation increases the observing agents' trust in the donor. Our results support these hypotheses and show that, apart from trust, responses to altruistic acts can have a rewarding or outcome-equalizing purpose. Our findings corroborate that the signaling benefits of altruistic acts that accrue in social exchange can ease the conditions for the evolution of social preferences.

1. Introduction

Humans frequently cooperate with non-kin others and incur costs to benefit them. The question of how such cooperative behavior can be explained has attracted considerable attention across several decades and disciplines (see West, El Mouden, & Gardner, 2011 for a critical review). A large body of literature has shown that cooperation can be a manifestation of self-interest if it is likely to be reciprocated with a benefit that outweighs its costs in the not-too-distant future (Trivers, 1971; Axelrod & Hamilton, 1981; Nowak & Sigmund, 1998). However, these explanations are restricted to interactions between members of relatively small groups, where cooperators and defectors can be identified and respectively targeted by reward or punishment (Bowles & Gintis, 2011: 63–70; Leimar & Hammerstein, 2001; Panchanathan & Boyd, 2003). Moreover, empirical evidence has accumulated suggesting that cooperative behavior may be motivated by psychological predispositions to benefit others (henceforth, social preferences) (Camerer, 2003: Ch. 2). However, since cooperative behavior is often costly, the evolution of social preferences in humans is difficult to explain in an individual-selectionist framework (although see Delton, Krasnow, Cosmides, & Tooby, 2011). This has led to a renewed interest in models of multi-level selection (Wilson, 1975; Gintis, 2000; Boyd, Gintis, Bowles, & Richerson, 2003).

Models of multi-level selection assume that there is both competition between individuals (within groups) and between groups, and groups with a higher proportion of cooperative individuals will be more likely to survive inter-group competition and adverse environmental conditions. For cooperation to be sustained in a population, positive assortment of cooperators, i.e., the higher likelihood of cooperators interacting with cooperators than with non-cooperators, must outweigh the ratio of costs c (for the cooperator) to benefits b (for the rest of the group) of cooperation (Eshel & Cavalli-Sforza, 1982; Bowles & Gintis, 2011: 52–59). However, since models of multi-level selection are mathematically equivalent to models of kin selection where genetic relatedness is implied by the limited dispersal of individuals within groups, some authors have argued that it is not necessary to resort to multi-level selection to explain the evolution of cooperation (West et al., 2011). We leave it to others to answer questions regarding to what degree population structures led to positive assortment of genetically related individuals in human prehistory and whether multi-level selection is necessary to explain how social preferences and cooperation have evolved. Instead, we

* Corresponding author. University of Oxford, Department of Sociology, Manor Road, Oxford OX1 3UQ, United Kingdom.
E-mail address: wojtek.przepiorka@sociology.ox.ac.uk (W. Przepiorka).

1090-5138/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.evolhumbehav.2012.11.005

argue that cooperative acts can be credible signals of an individual's social preferences and, through favorable treatment of these individuals in social exchange, ease the conditions for their evolution, whether in a kin or a multi-level selection framework.

1.1. Altruism as a signal of trustworthiness

Gintis, Smith, and Bowles (2001) show that cooperation can evolve as a costly signal of an unobservable but relevant quality, if this quality is causally related to an individual's ability to cooperate (see also Leimar, 1997; Roberts, 1998; Lotem, Fishman, & Stone, 2003; Smith & Bliege Bird, 2005). In the simplest case, there are high-quality and low-quality types who incur low costs (c_1) or high costs ($c_2 > c_1$), respectively, from sending the signal. If the benefits (s) from being interacted with, conditional on having sent the signal, compensate the high-quality types but not the low-quality types ($c_2 > s > c_1$), only the high-quality types can afford to send it and thus will be identified as such. If, moreover, sending the signal yields a higher net benefit for the sender than not sending the signal, type-separating behavior can evolve in which high-quality types send a signal, low-quality types do not send a signal, and agents are only interacted with if they sent a signal. Gintis et al. (2001) also analyze the evolutionary dynamics of their model and show that cooperation as a type-separating signal is evolutionarily stable under plausible conditions.

These predictions also hold if social preferences are the unobservable quality of interest. Agents with social preferences are the high-quality types, who derive a psychological reward ($r_1 > 0$) from benefiting others, whereas individuals lacking social preferences are the low-quality types, who only care about their own payoffs ($r_2 = 0$). Although the material costs are the same for both types ($c_2 = c_1 = c$), the psychological rewards make it "cheaper" for the high-quality type to cooperate ($c > c - r_1$). The condition that must hold for cooperation to be a type-separating signal is $c > s \geq 0$. In other words, high-quality types cannot be fully compensated in material terms for the costs they incur. In fact, their cooperative acts must be truly altruistic (henceforth, altruistic acts). This requires that the existence of individuals with social preferences is maintained by another evolutionary mechanism (e.g. kin and/or multi-level selection). However, as long as $s > 0$, cooperators receive partial compensation, which we call signaling benefits. We will show next that altruistic acts can induce signaling benefits through social exchange and that this can ease the conditions for the evolution of social preferences.

Social exchange among unrelated individuals has been an important part of human sociality for tens of thousands of years and arguably a driving force in the evolution of the human mind (Cosmides & Tooby, 1992). The upper half of Fig. 1B shows a Person X and a Person Y engaging in social exchange that is not based on a formally binding agreement (Dasgupta, 1988; Coleman, 1990: Ch. 5). The social exchange can be mutually beneficial if Person X makes a transfer x first and Person Y makes a back transfer y, which is tripled to reflect the gains from trade. While a selfish Person Y has a real incentive to keep x without sending back y, a Person Y with social preferences will make a back transfer y, such that $3y > x$. For Person X, a trust problem arises as he or she does not know whether Person Y is cooperative and will make a back transfer that is sufficiently high. Referring to the vast social science literature on social exchange (e.g. Ostrom & Walker, 2003; Fehr, 2009), we call Person Y's cooperative behavior trustworthiness and we call Person X's transfer, which is motivated by the expectation of gain from Person Y's back transfer, trust (see also definitions in bottom half of Fig. 1B).

Now suppose that Person X, before engaging in social exchange with Person Y, observes Person Y in the situation depicted in Fig. 1A. Here, Person Y has the opportunity to perform an altruistic act in the form of a charitable donation. Then, Person X can condition his or her transfer in the social exchange on whether Person Y acted altruistically (Y_1) or not (Y_2). Since only a Person Y with social preferences will both give to charity and make a back transfer in social exchange, Person X can infer Person Y's type from his or her donation to charity or the lack of it. Consequently, while Person Y_2 will be disregarded by Person X, Person Y_1 will be partly compensated for his or her altruistic act by the gains he or she makes from trade (i.e. $c' = c - s$, where c are the costs of the altruistic act and $s = x - y$ are the signaling benefits). In addition, since Person X benefits from the social exchange with a trustworthy interaction partner, the benefits for the group (excluding Person Y) increase as well (i.e. $b' = b + 3y - x$, where b are the group benefits from Person Y's altruistic act and $3y - x$ are Person X's gains). Now, as $c'/b' < c/b$, signaling eases the conditions for the evolution of social preferences, because it lowers the degree of positive assortment necessary to maintain such traits in the population.

This evolutionary argument implies that what we should observe today is that altruistic behavior and trustworthiness are correlated. Moreover, individuals acting altruistically will be trusted more in social exchange because they will be expected to have social preferences and thus to be trustworthy. Note that the reverse is not true in general. That is, observing someone behaving trustworthily does not necessarily tell us that this person has social preferences – he or she could be selfish and behave trustworthily to acquire a reputation for being trustworthy and to be trusted more in the future (Bolton, Katok, & Ockenfels, 2004). However, here we exclude this possibility with a one-shot (i.e. non-repeated) game design, in which a selfish Person Y has no incentive to act trustworthily and thus altruistic acts can be a signal of trustworthiness via social preferences only.

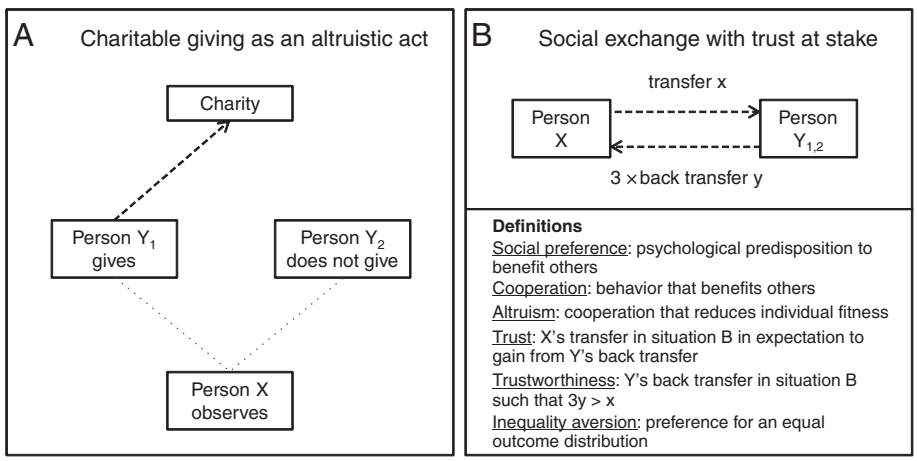

Fig. 1. Person X decides how much to transfer to Person Y in social exchange (B) contingent on Person Y's decision to give to charity or not (A).

1.2. Previous findings

There is ample evidence for a positive correlation between altruistic behavior and trustworthiness measured in laboratory experiments with economic games (Barclay, 2004; Ashraf, Bohnet, & Piankov, 2006; Chaudhuri & Gangadharan, 2007; Albert, Güth, Kirchler, & Maciejovsky, 2007; Blanco, Engelmann, & Normann, 2011; Fehrler unpublished; Gambetta & Przepiorka unpublished). Four of these experiments were also designed to investigate whether subjects who act altruistically are thereafter trusted more by third parties in social exchange, and they find support for this conjecture (Barclay, 2004; Albert et al., 2007; Fehrler unpublished; Gambetta & Przepiorka unpublished). Barclay and Willer (2007) and Sylwester and Roberts (2010) provide similar evidence from experiments with public good games. However, there is experimental evidence showing that subjects who help others or donate more to charity receive more in return from third parties (Wedekind & Milinski, 2000; Milinski, Semmann, & Krambeck, 2002). Thus, observing subjects in social exchange responding positively to altruistic acts does not tell us to what extent these responses reflect trust and to what extent they are mere transfers of resources intended to unconditionally reward the altruistic individual. Moreover, there is compelling experimental evidence that some subjects prefer egalitarian outcomes (Bolton & Ockenfels, 2000; Dawes, Fowler, Johnson, McElreath, & Smirnov, 2007; Fehr & Schmidt, 1999). Since altruism entails giving away resources or incurring costs, positive responses to altruistic acts could also be a manifestation of inequality aversion. However, while trust is motivated by pure self-interest, rewards or responses based on inequality aversion are not and would thus remain in need of an evolutionary explanation.

In our computerized laboratory experiment, we test whether altruistic behavior in the form of charitable giving is indeed correlated with trustworthiness. Moreover, we test whether a charitable donation increases the observing agents' trust in the donor. Our experimental design allows us to disentangle trust from rewarding and outcome-equalizing transfers as responses to altruistic acts.

2. Methods

Cox (2004) was the first to experimentally combine the dictator game (Forsythe, Horowitz, Savin, & Sefton, 1994) and the investment game (Berg, Dickhaut, & McCabe, 1995) to disentangle trustworthiness expectations from other motives behind trusters' decisions. He finds that trusters send higher amounts in the investment game than in the dictator game and attributes this difference to trusters' trustworthiness expectations. In our experiment, we take a similar approach. We give Person Y subjects the opportunity to donate part of their endowment to a charitable organization and disentangle the motives behind responses to these altruistic acts by using Person X subjects' transfers in the dictator game and the exchange game (a variant of the investment game).

2.1. Experimental games

Table 1 presents the dictator game (d) and the exchange game (e). In the dictator game, Person X and Person Y are endowed with G_X and G_Y Swiss francs (CHF), respectively. Next, Person X can decide to give up part or all of his or her endowment ($0 \le x_d \le G_X$) and transfer this amount to Person Y. The dictator game ends with Person X getting $G_X - x_d$ and Person Y getting $G_Y + x_d$. The exchange game extends the dictator game by giving Person Y the possibility to make a back transfer. That is, Person Y in the second mover position can decide to give up part of his or her amount ($0 \le y \le G_Y + x_e$) and transfer it to Person X. Unlike the transfer of Person X in both games, the amount transferred by Person Y is tripled. The exchange game ends with Person X getting $G_X - x_e + 3y$ and Person Y getting $G_Y + x_e - y$. Note that x_d and x_e denote Person X's transfer to Person Y in the dictator and exchange games, respectively. Given that, in the dictator game, Person Y does not have a possibility to make a back transfer, Person X's transfer x_d cannot be motivated by trustworthiness expectations. Moreover, if initial endowments are equal (i.e. $G_X = G_Y$), Person X's transfer in the dictator game cannot be motivated by inequality aversion either.

Table 1
Dictator and exchange game.

Dictator game (d)			Exchange game (e)		
Person X		Person Y	Person X		Person Y
G_X		G_Y	G_X		G_Y
$-x_d$	→	$+x_d$	$-x_e$	→	$+x_e$
$G_X - x_d$		$G_Y + x_d$	$G_X - x_e$		$G_Y + x_e$
			$+3y$	←	$-y$
			$G_X - x_e + 3y$		$G_Y + x_e - y$

2.2. Measuring trust and trustworthiness

Cox (2004) suggests that the difference between a transfer in the exchange game and in the dictator game measures trust because it nets out responses that are based on other motives, leaving the part of the exchange game transfer that is only based on trustworthiness expectations. However, this measure implies that the various motives additively affect subjects' transfer decisions. This assumption has been criticized on the grounds that the two games may put subjects in different mental frames (Fehr, 2009). It is plausible that the dictator game evokes more altruistic motives in subjects than the exchange game, with the effect that the transfer difference between the two games would underestimate trust.

In our study, we try to meet this objection in two ways. First, we balance the framing effects of the two games by presenting all decision situations to subjects on the same screen. This compels subjects to compare the different situations with each other. Second, we also measure trust in an alternative way. We regress the transfers in the exchange game on the expected back transfers, controlling for other motives by adding the dictator game transfer as a control variable. Then, we hold Person X subjects' dictator game transfers constant and measure their trust as the part of the exchange game transfer that can be attributed to their trustworthiness expectation only. This measure closely matches our definition of trust. Moreover, the regression model allows us to assess the extent to which other motives affect exchange game transfers. A coefficient estimate for dictator game transfers smaller than one would indicate that other motives affect transfers in the exchange game to a lesser extent than in the dictator game. In this case, the transfer difference alone would serve to underestimate trust.

We measure expected trustworthiness by asking Person X subjects what amount they expect each Person Y type to transfer back for hypothetical transfers of CHF 0, 8, and 16. We measure trustworthiness as a Person Y's back transfer conditional on a Person X's transfer.

2.3. Experimental design and procedure

To disentangle the motives behind responses to altruistic acts, we vary subjects' endowments, games, and the possibility to donate to a charitable organization in a $2(G_X = G_Y$ vs. $G_X > G_Y) \times 2$(dictator game vs. exchange game) $\times 2$(Y can donate vs. Y cannot donate) factorial, within-subject design. Upon arrival at the lab, subjects are randomly assigned to be a Person X or a Person Y and stay in their role throughout the experiment. In addition, Person Y subjects are randomly assigned to one of three conditions (see Fig. 2 below). In Condition 1, Person Y has the possibility of a one-time donation to a charitable organization. In conditions two and three, Person Y has no such possibility. In Condition 2, Person Y is endowed with the same

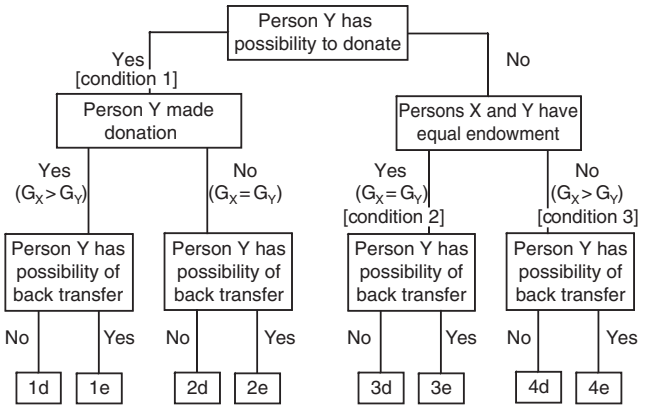

Fig. 2. Person X's decision situations (1d through 4e). The letters 'd' and 'e' stand for 'dictator game' and 'exchange game', respectively. The numbers 1 through 4 stand for different Person Y types. Subjects' endowments are equal (G = CHF 16) or Person Y has a lower endowment (G_Y = CHF 10), either after a charitable donation (Type 1) or by design (Type 4).

amount as Person X ($G_X = G_Y$ = CHF 16) and in Condition 3 Person Y's endowment is lower than Person X's endowment (G_X = CHF 16 > G_Y = CHF 10). At the beginning of the experiment, Person Y subjects in Condition 1 can decide whether or not to make a donation of CHF 6 to one of three organizations. They can choose from Amnesty International (AI), the International Committee of the Red Cross (ICRC), and Médecins Sans Frontières (MSF). Note that subjects who decide to donate are left with an endowment of CHF 10 or otherwise keep CHF 16. This corresponds to the endowments in conditions 3 and 2, respectively. Consequently, Person Y subjects differ with respect to the maximum amount they can send back to a Person X in the exchange game. In Condition 1, this difference is determined by the possibility to make a donation and in conditions 2 and 3this difference is determined by design. Subjects in the role of Person X face eight different decision situations. Fig. 2 presents the eight decision situations schematically.

Subjects make all possible decisions before they are randomly paired with another subject and payoffs are calculated and presented to them. The eight decision situations are presented to Person X subjects on one screen simultaneously. On the subsequent screen, we ask Person X subjects to state their expectations with respect to Person Y subjects' back transfers in the exchange game with hypothetical transfers of CHF 0, 8, and 16. Finally, Person Y subjects are asked to decide upon the amount they want to send back to Person X for every possible amount a Person X could transfer to them. The experimental procedure is described in more detail in the online supplement, available on the journal's website at www.ehbonline.org. www.ehbonline.org.

2.4. Hypotheses

Our first hypothesis is that donors to charity are more trustworthy than non-donors. We test our first hypothesis by regressing Person Y subjects' back transfers on an interaction of Person Y subjects' type (donor vs. non-donor) with Person X subjects' transfers.

Our second hypothesis is that Person X subjects expect donors to be more trustworthy than non-donors and therefore trust them more. Following the discussion in section 2.2, we test our second hypothesis in two ways. First, we calculate and compare the differences in actual exchange game and dictator game transfers to donors and non-donors (see Fig. 2 above). However, a comparison of these differences between donors and non-donors (1e–1d vs. 2e–2d) may be influenced by higher trust in non-donors due to non-donors' higher endowments (i.e. non-donors have more to send back). Therefore, we also compare transfer differences in the two games between donors and Person Y subjects without an option to donate and a low endowment (1e–1d vs. 4e–4d), as well as between non-donors and Person Y subjects without an option to donate and a high endowment (2e–2d vs. 3e–3d). Second, we regress Person X subjects' exchange game transfers on their transfers in the dictator game and their trustworthiness expectations towards each Person Y type. Based on this regression model estimation, we compute the differences in Person X subjects' exchange game transfers to each Person Y type that can only be attributed to differences in trustworthiness expectations.

Previous studies' findings suggest that the transfer decisions of Person X subjects may be caused by inequality aversion and/or a preference to reward altruistic acts. We expect to replicate these findings. That is, we expect to find higher transfers in the dictator game where Person Y has a lower endowment by design (inequality aversion: 4d>3d) and to find still higher transfers in the dictator game where Person Y has a lower endowment because he or she donated to charity (preference to reward altruistic acts: 1d>4d). Also, in accord with the results obtained in previous experiments with charitable giving (Albert et al., 2007; Milinski et al., 2002), we expect to find higher transfers to donors than to non-donors in both the dictator (1d>2d) and the exchange game (1e>2e).

3. Results

3.1. Charitable giving and trustworthiness

Of the 42 Person X subjects who had the opportunity to make a donation, 26 (62%) chose to do so. Fig. 3 shows Person Y back transfers at Person X transfer levels of CHF 0, 8 and 16. The joint hypotheses test of back transfer differences between donors and non-donors at all 17 transfer levels (0 through CHF 16: $F_{17,41} = 2.04$, $p = 0.032$) indicates that donors send back significantly higher amounts than non-donors (see Table A2 in the online supplement, available on the journal's website at www.ehbonline.org). Moreover, the slope coefficients for donors and non-donors in the regression of back transfers on transfers are significantly different at the 10% level ($t = 1.71$, $p = 0.094$). This shows that donors reciprocate higher transfers with higher back transfers than non-donors (see Table A3 in the online supplement, available on the journal's website at www.ehbonline.org).

3.2. Trust I

Our first measure of trust is the difference between exchange game and dictator game transfers. Fig. 4 shows Person X subjects' average

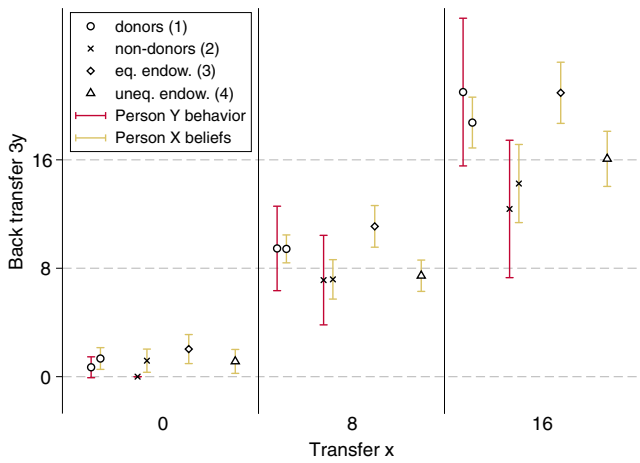

Fig. 3. Person X beliefs about Person Y back transfers and Person Y actual back transfers conditional on Person X transfers of CHF 0, 8 and 16. Person Y back transfers conditional on all 17 transfer levels are listed in Table A2 in the online supplement , available on the journal's website at www.ehbonline.org.

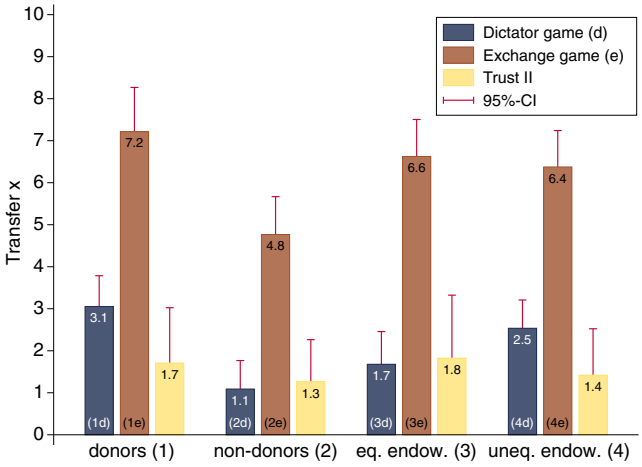

Fig. 4. Person X mean transfers in the eight decision situations (1d through 4e) and predicted transfers. The letters 'd' and 'e' stand for 'dictator game' and 'exchange game', respectively. The numbers 1 through 4 stand for different Person Y types.

transfers in the eight different decision situations (the first two bars in each three-bar grouping) and Table 2 lists the transfer differences discussed in the following. In accord with previous studies' findings, we observe a substantial and statistically significant difference between exchange game transfers to donors and non-donors (Table 2, row 1). However, since dictator game transfers between donors and non-donors differ to a similar extent (Table 2, row 2), we cannot be certain whether the difference in exchange game transfers can be attributed to a difference in trustworthiness expectations or other motives. Hence, we first compare the difference of dictator game and exchange game transfers between donors and non-donors (Table 2, row 3).

This measure of trust yields a positive but statistically insignificant difference between donors and non-donors. Comparing the transfer differences of donors and Person Y subjects without the option to donate and a low endowment yields a similar result. The difference is positive but also statistically insignificant (Table 2, row 4). However, when we compare the transfer differences of non-donors and Person Y subjects without the option to donate and a high endowment, we find a substantial and statistically significant negative difference (Table 2, row 5).

According to the discussion in section 2.2, the difference between exchange game and dictator game transfers alone may underestimate trust because other motives may affect exchange game transfers to a smaller extent than dictator game transfers. Moreover, looking at a direct measure of trustworthiness expectations may be more informative. If Person X subjects' expectations about donors' and non-donors' trustworthiness do not differ as predicted by our second hypothesis, then arguments based on costly signaling can be ruled out.

Table 2
Transfer differences.

			$F_{1,55}$	p
1	1e–2e	2.45	20.03	<0.001
2	1d–2d	1.96	75.02	<0.001
3	1(e–d)–2(e–d)	0.48	0.88	0.353
4	1(e–d)–4(e–d)	0.32	0.65	0.425
5	2(e–d)–3(e–d)	−1.27	14.72	<0.001
6	4d–3d	0.86	13.68	0.001
7	1d–4d	0.52	7.39	0.009
8	2d–3d	−0.59	10.79	0.002

Notes: The numbers 1d through 4e in the second column denote Person X transfers in each of the eight decision situations (see Fig. 2). The Wald tests of simple and composite linear hypotheses are based on the OLS regression model presented in Table A6 in the online supplement, available on the journal's website at www.ehbonline.org.

3.3. Trust II

Fig. 3 above also shows the expected back transfers as stated by Person X subjects at hypothetical transfer levels of CHF 0, 8 and 16 and for each Person Y type. A clear picture is given by the joint hypotheses test of the differences in Person X expectations regarding each Person Y type at the three transfer levels. Person X subjects expect significantly higher back transfers from donors than from non-donors ($F_{3,55}=4.45$, $p=0.007$) and they also expect significantly higher back transfers from donors than from unequally endowed Person Y subjects without a possibility to donate ($F_{3,55}=3.74$, $p=0.016$). Person X subjects expect to receive most back from equally endowed Person Y subjects without a possibility to donate (see Table A4 in the online supplement, available on the journal's website at www.ehbonline.org). Moreover, regressing expected back transfers on hypothetical transfers yields a significantly larger slope coefficient for donors than for non-donors ($F_{1,55}=12.56$, $p<0.001$) and for donors than for unequally endowed Person Y subjects without a possibility to donate ($F_{1,55}=5.06$, $p=0.029$). This indicates that subjects expect higher transfers to be reciprocated by higher back transfers from donors and provides evidence for motives based on trustworthiness expectations in the exchange game (see Table A5 in the online supplement, available on the journal's website at www.ehbonline.org). But do Person X subjects act on their trustworthiness expectations?

To answer this question, we regress Person X subjects' transfers in the exchange game on their transfers in the dictator game and their trustworthiness expectations about the four Person Y types (see Model M1 in Table A7 in the online supplement, available on the journal's website at www.ehbonline.org). First of all, the coefficient estimate for dictator game transfers is 0.616 and significantly smaller than one ($F_{1,55}=34.17$, $p<0.001$). This indicates that other motives affect transfers in the exchange game to a lesser extent than in the dictator game and that the difference between exchange game and dictator game transfers is an overly conservative measure of trust. Therefore, based on this model estimation, we compute the part of Person X subjects' exchange game transfers that can only be attributed to their trustworthiness expectations. In Fig. 4, the third bar in each three-bar group shows this measure of trust towards the four Person Y types. These figures show that, on average, Person X subjects transfer CHF 0.45 more to donors than to non-donors *because* they expect donors to be more trustworthy than non-donors ($z=2.79$, $p=0.005$). For the same reason, Person X subjects transfer on average CHF 0.29 more to donors than to unequally endowed Person Y subjects without a possibility to donate ($z=2.53$, $p=0.011$). The largest difference in exchange game transfers that can be attributed to the difference in trustworthiness expectations is between non-donors and equally endowed Person Y subjects without a possibility to donate and amounts to CHF 0.56 ($z=2.21$, $p<0.027$). These results support our second hypothesis.

3.4. Other motives

Finally, our study replicates previous findings. First, the amounts transferred in the dictator game with Person Y having a lower endowment are significantly higher, indicating that inequality aversion is important (Table 2, row 6). Second, the fact that, in the dictator game, donors to charity receive higher transfers than subjects without an opportunity to donate suggests that some subjects have a preference to reward altruistic acts (Table 2, row 7). Our evidence also suggests that non-donors are punished (Table 2, row 8).

4. Discussion

Empirical evidence suggests that human cooperation can be motivated by social preferences, and the evolution of social

preferences can be explained with kin or multi-level selection models. However, it has been shown that cooperation can also evolve as a costly signal of an unobservable but relevant quality, if this quality is causally related to an individual's ability to cooperate. We propose that if a proportion of individuals with social preferences is maintained in the population through kin or multi-level selection, cooperative acts that are truly altruistic can signal trustworthiness, and the signaling benefits that accrue in social exchange can ease the conditions for the evolution of social preferences. In social exchange, trust problems arise as an actor does not know whether his or her potential exchange partners are cooperative or not. However, since a person with social preferences will both engage in altruistic behavior and be cooperative in social exchange, the actor can infer his or her potential partners' types from their altruistic behavior and choose a trustworthy partner accordingly. Then, the gains from trade partly compensate the altruistic individuals for their altruistic acts and increase overall group benefits through the actor's gains. This eases the conditions for the evolution of social preferences, because it lowers the degree of positive assortment necessary to maintain such traits in the population.

This account of the evolution of altruistic behavior in humans implies that what we should observe today is that, first, altruistic behavior and trustworthiness are correlated and, second, altruists are expected to be more trustworthy and therefore are trusted more in social exchange. However, observing agents' positive responses to altruistic acts in social exchange does not tell us to what extent these responses reflect trust and to what extent they are only rewarding or outcome-equalizing transfers of resources. Our experimental design allows us to isolate trust from these other responses.

Our evidence is consistent with our hypotheses. First, we find that donors to charity, despite the fact that they have less to transfer back, transfer back significantly higher amounts in social exchange than non-donors. Second, we find that, in social exchange, subjects transfer significantly higher amounts to donors than to non-donors *because* they expect to receive back more from donors than from non-donors. We also find evidence that subjects reward donors and punish non-donors in both the dictator and exchange games. Moreover, we find evidence for inequality aversion and endowment effects. Those who have a lower endowment by design receive more in the dictator game than those who have an equal endowment, but the latter are trusted more in social exchange.

An alternative explanation for our findings could be constructed by combining an indirect reciprocity mechanism that explains the evolution of cooperative strategies with an argument of maladaptation to the game structure of our experiment. Panchanathan and Boyd (2004) show that if a public good game is followed by an infinitely repeated indirect reciprocity game, a strategy that contributes to the public good and thereafter refuses to help free-riders, but helps other contributors in the indirect reciprocity game, can stabilize cooperation in the public good game. Applied to the game structure in our experiment (charitable giving followed by a one-shot exchange game) it is obvious that such a "shunner" strategy would not be evolutionarily stable because it forgoes fitness-enhancing benefits by giving away resources in the last move of the game. However, recently Delton et al. (2011) have convincingly argued that cooperation in one-shot games could be a maladaptation. Since the costs of mistaking a repeated interaction for a one-shot interaction are so much larger than the costs of mistaking a one-shot interaction for a repeated interaction, evolution might have led to motivational and representational systems in the human brain that are specialized in avoiding the first type of error while accepting occasional losses due to the second type of error.

We acknowledge that the behavior we observe could also be explained by indirect reciprocity theory *cum* maladaptation. Nevertheless, we find our account more plausible because there is ample evidence from lab experiments that subjects are very capable of distinguishing one-shot from repeated games and act accordingly when playing them (Keser & van Winden, 2000; Gächter & Falk, 2002). However, it is not our aim to discard other explanations of the evolution of cooperation. Instead, we believe that kin or multi-level selection plus signaling might have complemented mechanisms based on direct and indirect reciprocity in scenarios where repeated interactions and the accurate transmission of information about reputation were unlikely (Roberts & Sherratt, 2007; Bowles & Gintis, 2011: Ch. 4).

Supplementary Materials

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.evolhumbehav.2012.11.005.

Acknowledgments

We thank Michèle Belot, Oliver Curry, Andreas Diekmann, Charles Efferson, Claire El Mouden, Guillaume Fréchette, Katharina Michaelowa, David Myatt, the participants of the Nuffield College Postdoc Seminar at the University of Oxford and the CESS internal seminar at New York University, and two anonymous reviewers for their very helpful comments and suggestions. We are also grateful to Stefan Wehrli and Silvana Jud from DeSciL, the experimental laboratory at ETH Zurich, for their support with the experiment. This research was partly supported by the Swiss National Science Foundation (grant number 100017_124877).

References

Albert, M., Güth, W., Kirchler, E., & Maciejovsky, B. (2007). Are we nice(r) to nice(r) people? An experimental analysis. *Experimental Economics*, *10*, 53–69.
Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, *9*, 193–208.
Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*, 1390–1396.
Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons". *Evolution and Human Behavior*, *25*, 209–220.
Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society London B*, *274*, 749–753.
Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*, 122–142.
Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, *72*, 321–338.
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., & Ockenfels, A. (2000). ERC: a theory of equity, reciprocity, and competition. *American Economic Review*, *90*, 166–193.
Bowles, S., & Gintis, H. (2011). *A cooperative species: human reciprocity and its evolution.* Princeton: Princeton University Press.
Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences USA*, *100*, 3531–3535.
Camerer, C. F. (2003). *Behavioral game theory.* Princeton, NJ: Princeton University Press.
Chaudhuri, A., & Gangadharan, L. (2007). An experimental analysis of trust and trustworthiness. *Southern Economic Journal*, *73*, 959–985.
Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, MA: The Belknap Press of Harvard University Press.
Cosmides, L., & Tooby, J. (1992). Cognitive Adaptations for Social Exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The Adapted Mind* (pp. 163–228). New York: Oxford University Press.
Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*, 260–281.
Dasgupta, P. (1988). Trust as a commodity. In D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 49–72). Oxford: Basil Blackwell.
Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, *446*, 794–796.
Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences USA*, *108*, 13335–13340.
Eshel, I., & Cavalli-Sforza, L. L. (1982). Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences USA*, *79*, 1331–1335.
Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association*, *7*, 235–266.
Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*, 817–868.
Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*, 347–369.

Gächter, S., & Falk, A. (2002). Reputation and reciprocity: Consequences for the labour relation. *The Scandinavian Journal of Economics*, *104*, 1–26.
Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*, 169–179.
Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology*, *213*, 103–119.
Keser, C., & van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *The Scandinavian Journal of Economics*, *102*, 23–39.
Leimar, O. (1997). Reciprocity and communication of partner quality. *Proceedings of the Royal Society London B*, *264*, 1209–1215.
Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London B*, *268*, 745–753.
Lotem, A., Fishman, M. A., & Stone, L. (2003). From reciprocity to unconditional altruism through signalling benefits. *Proceedings of the Royal Society London B*, *270*, 199–205.
Milinski, M., Semmann, D., & Krambeck, H. -J. (2002). Donors to charity gain in both indirect reciprocity and political reputation. *Proceedings of the Royal Society London B*, *269*, 881–883.
Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by Image scoring. *Nature*, *393*, 573–577.
Ostrom, E., & Walker, J. (2003). *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*. New York: Russell Sage.
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.
Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *434*, 499–502.
Roberts, G. (1998). Competitive altruism: from reciprocity to the handicap principle. *Proceedings of the Royal Society London B*, *265*, 427–431.
Roberts, G., & Sherratt, T. N. (2007). Cooperative reading: Some suggestions for integration of the cooperation literature. *Behavioural Processes*, *76*, 126–130.
Smith, E. A., & Bliege Bird, R. (2005). Costly signaling and cooperative behavior. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life* (pp. 115–148). Cambridge, MA: MIT Press.
Sylwester, K., & Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters*, *6*, 659–662.
Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*, 35–57.
Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, *288*, 850–852.
West, S. A., El Mouden, C., & Gardner, A. (2011). Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*, *32*, 231–262.
Wilson, D. S. (1975). A theory of group selection. *Proceedings of the National Academy of Sciences USA*, *72*, 143–146.