PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2010) **277**, 1315–1323
doi:10.1098/rspb.2009.1900
Published online 23 December 2009

Eyes are on us, but nobody cares: are eye cues relevant for strong reciprocity?

Ernst Fehr and Frédéric Schneider*

Institute for Empirical Research in Economics, University of Zürich, Blümlisalpstrasse 10, 8057 Zürich, Switzerland

Strong reciprocity is characterized by the willingness to altruistically reward cooperative acts and to altruistically punish norm-violating, defecting behaviours. Recent evidence suggests that subtle reputation cues, such as eyes staring at subjects during their choices, may enhance prosocial behaviour. Thus, in principle, strong reciprocity could also be affected by eye cues. We investigate the impact of eye cues on trustees' altruistic behaviour in a trust game and find zero effect. Neither the subjects who are classified as prosocial nor the subjects who are classified as selfish respond to these cues. In sharp contrast to the irrelevance of subtle reputation cues for strong reciprocity, we find a large effect of explicit, pecuniary reputation incentives on the trustees' prosociality. Trustees who can acquire a good reputation that benefits them in future interactions honour trust much more than trustees who cannot build a good reputation. These results cast doubt on hypotheses suggesting that strong reciprocity is easily malleable by implicit reputation cues not backed by explicit reputation incentives.

Keywords: altruism; strong reciprocity; trust game; cues; reputation

1. INTRODUCTION

Human altruism represents a huge outlier in the animal world (Boyd & Richerson 2005). Humans often behave altruistically towards genetically unrelated strangers, even if the chance of meeting these strangers again is extremely small and reputational concerns are unlikely to play a role (e.g. tipping an unknown taxi driver in a large foreign city). Altruistic behaviours in the absence of any opportunity of repeated interaction and reputation formation have been repeatedly shown in tightly controlled economic experiments (Camerer 2003; Fehr & Fischbacher 2003; Gintis *et al.* 2003). Experimental evidence (Fehr *et al.* 2002), social preference theories (Rabin 1993; Fehr & Schmidt 1999; Dufwenberg & Kirchsteiger 2004; Falk & Fischbacher 2006) and evolutionary theories (Gintis 2000; Henrich & Boyd 2001; Boyd *et al.* 2003; Bowles & Gintis 2004) also indicate that a special type of altruistic behaviour—strong reciprocity—plays a particularly important role in establishing and sustaining cooperation among strangers. Strong reciprocity is characterized by the willingness to altruistically reward cooperative acts and to altruistically punish norm-violating, defecting behaviours. As a consequence, strong reciprocity generates important incentives for cooperation among strangers.

The fact that important altruistic behaviours exist even in the absence of reputation incentives does not mean that such incentives are irrelevant. A large literature shows that human cooperation and other forms of prosocial behaviour are positively affected by the possibility of acquiring a 'good' reputation that may pay off in future interactions (Gächter & Fehr 1999; Wedekind & Milinski 2000; Milinski *et al.* 2001, 2002; Brown *et al.* 2004; Rege & Telle 2004; Rockenbach & Milinski 2006; Kurzban *et al.* 2007; Engelmann & Fischbacher 2009; Fehr *et al.* 2009). However, recent articles seem to suggest that much of human altruistic behaviour may merely be a response to subtle reputation cues that are not in fact related to the possibility of benefiting in future interactions from current altruistic acts (Haley & Fessler 2005; Bateson *et al.* 2006; Hagen & Hammerstein 2006; Burnham & Hare 2007; Rigdon *et al.* 2009). Haley and Fessler argue that reputation incentives in the ancestral evolutionary environment thoroughly moulded human social interactions because 'natural selection can be expected to have shaped human psychology to be exquisitely sensitive to cues that are (or were, under ancestral conditions) informative with respect to the likely profitability of co-operation in a given situation' (Haley & Fessler 2005, p. 249). These authors thus implemented a visual cue in an anonymous experimental game—eyes staring at the subjects during decision-making—a cue 'that, over the course of human evolution, would have reliably indicated the potential observability of one's behaviour' (p. 249). Haley & Fessler (2005), Bateson *et al.* (2006), Burnham & Hare (2007) and Rigdon *et al.* (2009) indeed found that eyes staring at the subjects cause an increase in prosocial behaviours in anonymous games such as the dictator game.

In this paper we examine whether a reputation cue like that implemented in Haley & Fessler (2005) also affects strong reciprocity, by implementing an anonymous, one-shot trust game in which a trustor can send money to a trustee; the experimenter then quadruples this amount, so that the trustee receives four times the amount sent. The trustee observes how much the trustor has sent and can then send back as little or as much money as he wants. Thus, the trustee can altruistically reward trustors who have sent money, which constitutes an instance of

* Author for correspondence (frederic@iew.uzh.ch).

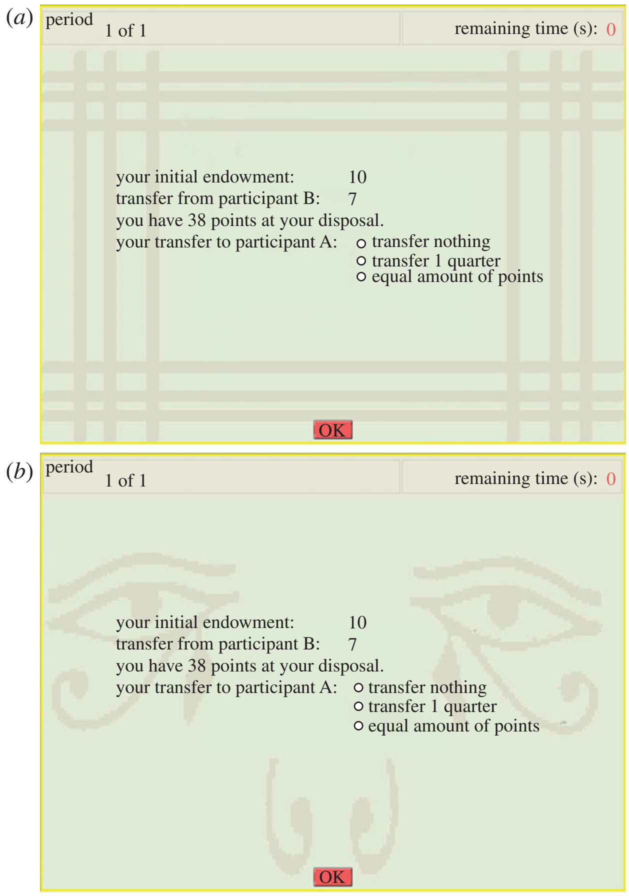
Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2009.1900 or via http://rspb.royalsocietypublishing.org.

Received 3 November 2009
Accepted 4 December 2009

strong reciprocity. By comparing the eye cue condition with a baseline condition without such cues we can assess the impact of eye cues on strong reciprocity.

In addition to the eye cue condition we implement another reputation condition in which subjects face a real pecuniary incentive for acquiring a good reputation. Previous work has argued that eye cues activate reputational concerns, but has not explicitly compared the effect of eye cues with the effect of explicit pecuniary reputation incentives. If humans are indeed 'exquisitely sensitive' to reputation cues even if they carry no real pecuniary incentive power, subjects should generate patterns in the eye cue condition that resemble the effects of explicit pecuniary reputation incentives. Our design enables us to conduct this comparison and investigate the relative importance of reputation cues for altruistic behaviour.

We also go beyond previous work by examining which—if any—subjects respond to the implicit reputation cue, because we measure subjects' degree of selfishness and opportunism with a Machiavellianism questionnaire (Christie & Geis 1970). Assessing individual differences in subjects' responses to reputation cues is important because on average one might find a null effect that hides important inter-individual differences. Implicit reputation cues could increase the altruistic behaviour of prosocial subjects (i.e. those who score low on the Machiavellianism (Mach) scale). This has important consequences on the interpretation of altruistic behaviours in anonymous one-shot experiments. If prosocial subjects primarily respond to the implicit reputation cue, it is possible to argue that they are mostly prone to all kinds of *other* subtle reputation cues that are often not controlled for by the experimenter in the typical laboratory experiment (e.g. the mere presence of other subjects and the experimenter in the room, or simply hearing human voices). It would then be more plausible to attribute the baseline altruism observed in anonymous one-shot experiments to such uncontrolled reputation cues. However, if prosocial subjects do not respond to the eye cues, it is hard to argue along these lines. It is then implausible to attribute the observed altruistic behaviours in anonymous one-shot experiments to uncontrolled subtle reputation cues. Thus, by measuring subjects' Mach scores we can put important constraints on the interpretation of altruistic behaviours in anonymous one-shot experiments.



(*a*)

period 1 of 1 remaining time (s): 0

your initial endowment: 10
transfer from participant B: 7
you have 38 points at your disposal.
your transfer to participant A: ○ transfer nothing
○ transfer 1 quarter
○ equal amount of points

OK

(*b*)

period 1 of 1 remaining time (s): 0

your initial endowment: 10
transfer from participant B: 7
you have 38 points at your disposal.
your transfer to participant A: ○ transfer nothing
○ transfer 1 quarter
○ equal amount of points

OK

Figure 1. Trustee decision screen. (*a*) Baseline background. (*b*) Eyespots background.

2. THE EXPERIMENT

(a) *Experimental procedure*

We measured strong reciprocity as second-mover behaviour in a series of one-shot trust games. A trustor and a trustee interact with each other in a trust game; the trustor can send money to the trustee, which is then multiplied by the experimenter so that the overall money available to the two parties increases. The trustee can then send back none or some of the money to the trustor. Fairness norms typically demand that the trustee sends back some of the money he received, but the trustee is completely free to send back nothing if he likes. Details of the trust game are described in §2*b* below.

Our experimental design includes three treatments: a baseline treatment where the trustee faces a neutral background screen (figure 1*a*); an 'implicit reputation cue' treatment where the background screen features eyespot-like shapes (figure 1*b*) similar to those in Haley & Fessler (2005); and an 'explicit reputation' treatment where the current trustor is informed of the trustee's decisions in the previous periods (same background screen as in the baseline treatment).

Subjects were seated in isolated compartments and were assigned either to the role of a trustor or a trustee. They maintained their roles throughout the experiment. They then played 10 periods of the trust game, with randomly re-matched partners each period.

We conducted eight sessions (three sessions each in baseline and implicit, two in explicit treatment), each session involving 36 participants, hence 288 participants in total (144 trustors and 144 trustees). The decisions in a group of subjects who interact with each other over the 10 periods are statistically not independent. In order to establish statistically independent observations, we created three matching groups per session, each consisting of 12 subjects. Only the subjects within a matching group were matched with each other during the experiment, generating three independent observations per session. With three matching groups per session, we have nine independent observations both in the baseline treatment and the implicit cues treatment, and six independent observations in the explicit reputation treatment.

Immediately after the end of the last period, the participants had to fill out a questionnaire (emotional state, fairness attitudes, Machiavellianism, trust, socioeconomic

Table 1. This table shows the monetary payoffs that are associated with each action combination. The first number in each cell denotes the trustor's payoff while the second number denotes the trustee's payoff. For example, if the trustor invests 7 points and the trustee 'compensates', the trustor earns 10 points and the trustee earns 31 points.

	trustee					
trustor	nothing		compensate		equalize	
1 point	9	14	10	13	11.5	11.5
4 points	6	26	10	22	16	16
7 points	3	38	10	31	20.5	20.5
10 points	0	50	10	40	25	25

data). After completion, participants were paid a show-up fee of CHF 10 plus their earnings from the experiment, at the rate of 1 point = CHF 0.2. In total, a session lasted approximately 2 h and subjects earned on an average CHF 48.88 ($41.77).

(b) *Game design*

Each period of the experiment was a one-shot trust game. At the beginning of each period trustors and trustees were endowed with 10 points. The game itself consisted of two stages: an investment stage, where trustors had to decide how many points they would transfer to their current trustee, and a back-transfer stage, where trustees had to decide how much they wanted to back-transfer to the trustor. The amount trustors invested was quadrupled and transferred to the trustee. Trustors could choose between four possible transfers: 1 point, 4 points, 7 points, or 10 points. Trustees had three options: they could back-transfer either nothing, or the amount the trustor sent (henceforth 'compensate'), or they could back-transfer an amount that equalized the period payoff between trustor and trustee (henceforth 'equalize'). When the trustee determined the back-transfer he was perfectly informed about the trustor's choice and thus did not have to form beliefs about the size of the investment.

In table 1 we show the payoff matrix that corresponds to our trust game. The first number in each cell of the matrix represents the trustor's payoff, the second number denotes the trustee's payoff. The matrix shows that for any given investment level, the trustee is always best off in terms of monetary payoff by back-transferring nothing. This means that positive back-transfers (i.e. the choices 'compensate' and 'equalize') can be regarded as altruistic acts because the trustee gives up some of his own payoff to increase the trustor's payoff.

(c) *Personality questionnaires and fairness norms*

As the response to the different treatments may be heterogeneous depending on the subject's degree of selfishness, we also measured each subject's Mach score. For this purpose we used the MACH-IV Machiavellianism Questionnaire (Christie & Geis 1970), which provides a measure of selfishness and opportunism. Recent results from a neuroeconomic study (Spitzer *et al.* 2007) indicate that Machiavellian subjects are much less willing to share money in a dictator game, and respond much more strongly to pecuniary punishment threats for norm violations. Moreover, subjects' Machiavellianism also correlated strongly with activation in the lateral orbitofrontal cortex that is known to be reliably activated when subjects face punishing stimuli. Thus, behavioural and neurophysiological evidence suggests that subjects' Machiavellianism may affect their responses to our treatment conditions.

We also measured subjects' fairness standards by asking them the following question: 'suppose that participant A (i.e. the trustor) transferred 10 points to participant B (i.e. the trustee). B then chose "compensate". How fairly do you judge this behaviour?' (Additions in brackets did not appear in the questionnaire.) Subjects indicated their answer to this question on a Likert scale, coded from 1 ('very unfair') to 7 ('very fair'). Note that subjects with high fairness standards perceive the choice as unfair and therefore assign a low score to this question, while subjects with low fairness standards perceive the choice as fair and assign a high score.

3. HYPOTHESES

The implicit cues treatment measures the impact of implicit reputation cues on trustees' altruistic behaviour in the trust game. The explicit reputation treatment enables us to assess the effect of explicit pecuniary reputation incentives on trustees' behaviour. Thus, we can gain insight into the relative importance of the two kinds of reputation effects by comparing the effect of implicit cues with the effect of explicit reputation incentives.

Consider the baseline and the implicit reputation treatment condition first. The game played in these two conditions constitutes a true one-shot game because the players remain fully anonymous and they meet a different anonymous partner in each period. Therefore, if both players are completely selfish and want to maximize their money earnings, and the trustor knows this, the following outcome is predicted. The selfish trustee will always choose 'nothing' (i.e. his back-transfer is zero), and the trustor will invest the lowest possible amount, because he knows that the trustee will back-transfer nothing in any case.

However, there is a large literature indicating that a substantial share of experimental subjects is not completely selfish (see Fehr & Fischbacher 2003 for a review). This literature indicates that subjects may also have social motives such as inequity aversion (Fehr & Schmidt 1999; Dawes *et al.* 2007) or intention-based reciprocity (Rabin 1993; Dufwenberg & Kirchsteiger 2004; Falk & Fischbacher 2006). Inequity-averse trustees will choose the 'equalize' option, while trustees who interpret high investments as particularly kind acts will make more generous back-transfers in response to high investments. We summarize both these behaviours under the term 'altruistic rewarding' because they imply a benefit for the trustor at the expense of the trustee and they reward the trustor's cooperative investment.

A key question then is whether subjects' social preferences are affected by implicit reputation cues such as eyespots. Recent evidence (Haley & Fessler 2005; Bateson *et al.* 2006; Burnham & Hare 2007) suggests that eye cues affect prosocial behaviour in dictator games and public good games. In view of this literature, one would expect the trustees to respond to the eyespots in the implicit reputation treatment by making significantly higher back-transfers compared with the baseline treatment.

As we are interested in the impact of the implicit reputation cue on the trustees' social preferences, the fact that the trustee chooses his back-transfer with the exact knowledge of how much the trustor invested is important. This feature of our design ensures that unknown beliefs about the trustors' investments do not affect the trustees' choices. In this respect, our design differs substantially from the public goods experiments of Bateson *et al.* (2006) and Burnham & Hare (2007), because it is not clear why subjects change their contributions in response to a cue in a public goods experiment. In principle, the cue could cause a more optimistic belief about the other players' public good contributions, which will then lead to an increase in the subject's own contribution; it is known that many subjects are conditional cooperators (Fischbacher *et al.* 2001; Kurzban & Houser 2005; Croson 2007; Kocher *et al.* 2008)—that is, they are willing to contribute more to the public good if they believe that other group members contribute more. Alternatively, the reputation cue could have a direct impact on subjects' social preferences, implying that subjects are willing to contribute more for any given belief level. If the first hypothesis holds, the reputation cue does not affect subjects' social preferences; it 'only' renders their beliefs about others more optimistic, which then causes the change in behaviour. If the second hypothesis holds, the reputation cue has a direct effect on subjects' social preferences. In our experimental design a change in the trustees' behaviour cannot be attributed to changes in their beliefs because the trustees know the exact investment when they make their back-transfer. Thus, we can measure the impact of the implicit reputation cue for any given transfer level, which provides a clean behavioural measure of a change in social preferences.

Because we measure subjects' degree of Machiavellianism and their fairness standards, we are able to examine whether subjects who score differently on these measures respond differently in the different treatments. We expect, in particular, that highly Machiavellian subjects tend to back-transfer less in the baseline condition. It is also important to examine the impact of the implicit cue condition for subjects who score high and low on the Mach score. In particular, if the non-selfish subjects (i.e. those scoring low on the Mach score) are particularly responsive to the implicit reputation cue, one may be more inclined to attribute the observed prosociality in anonymous one-shot experiments to uncontrolled implicit reputational features. In contrast, if subjects' Mach scores do not affect the response to the implicit cue, one may have more confidence in the hypothesis that the prosocial behaviour in anonymous experiments is a true expression of subjects' social preferences and not just an artefact of uncontrolled implicit reputation cues.

In the explicit reputation treatment, the players' personal identities are still kept anonymous but we render the history of the trustees' back-transfers observable for their current trustors. Thus, each trustor can assess the past willingness of the current trustee to back-transfer resources. Because the trustees know this, even selfish trustees now have an incentive to choose 'compensate' or 'equalize', because in this way they can increase the likelihood that the trustors they face in the future (and know their past choices) will make large investments. This explicit reputation incentive ceases in the final period (when there will be no future encounters with trustors): the selfish trustees will defect in the last period, and only the trustees with social preferences will make positive back-transfers.

The effectiveness of the explicit reputation incentive requires that the trustees understand that their current back-transfers will affect average investments of future trustors. Thus, the explicit reputation incentive will only raise the trustees' back-transfers if the trustees exhibit this kind of rationality. Reputation incentives can also increase back-transfers of subjects with social preferences. They may, for example, choose 'equalize' instead of only 'compensate' when the pecuniary incentive coincides with their social motive. The hypothesis that explicit reputation incentives increase trustees' transfers is also backed by previous findings (Gächter & Falk 2002; Cochard *et al.* 2004).

Our measure of Machiavellianism enables us to examine whether there is a meaningful heterogeneity in trustees' responses to the explicit reputation incentive. In view of the behavioural and neurophysiological evidence documented in Spitzer *et al.* (2007), it seems plausible to conjecture that highly Machiavellian subjects respond more strongly to the explicit reputation incentive. Future trustors are likely to punish low back-transfers by lowering their investments. By definition, highly Machiavellian subjects are particularly susceptible to such threats. Therefore, they should respond more strongly to the pecuniary reputation incentives.

4. DOES THE IMPLICIT REPUTATION CUE MATTER?

In this section, we examine the impact of implicit reputation cues on trustees' back-transfers. If the implicit cue raises reputational concerns, the trustees in the implicit cue condition should make higher back-transfers than those in the baseline condition. Moreover, if the implicit cue has a sufficiently strong effect, the back-transfer pattern in the implicit cue condition should resemble the pattern in the explicit reputation condition. Finally, if the implicit reputation cue raises the trustees' back-transfers, this may also increase the trustors' investments because higher investments increase the trustors' payoffs if a sufficiently high share of trustees choose to equalize payoffs (see the final column in table 1).

Table 2 provides a first indication of the impact of the implicit cue condition. In the baseline condition, the average back-transfer is 6.28 points and the trustees' modal choice is 'nothing'. The average back-transfer in the implicit cue condition is even somewhat lower and the modal choice is also 'nothing'. The small difference in the means across conditions is not significant (Mann–Whitney test: $p = 0.402$, $n = 18$). The trustors' investment choices are also very similar across the two conditions. The median investment level in both conditions is 7; trustors in the baseline condition invest an average of 5.88, while the average investment in the implicit cue condition is 5.74 (Mann–Whitney test: $p = 0.825$, $n = 18$).

Table 2. Descriptive statistics of average, median and modal behaviour across treatments.

statistic	baseline	implicit	explicit
average back-transfer	6.28	5.36	13.86
median back-transfer	compensate	nothing	equalize
mode trustee decision	nothing	nothing	equalize
average investment	5.88	5.74	7.73
median investment	7	7	10
mode investment	10	10	10
number subjects	108	108	72
number matching groups	9	9	6

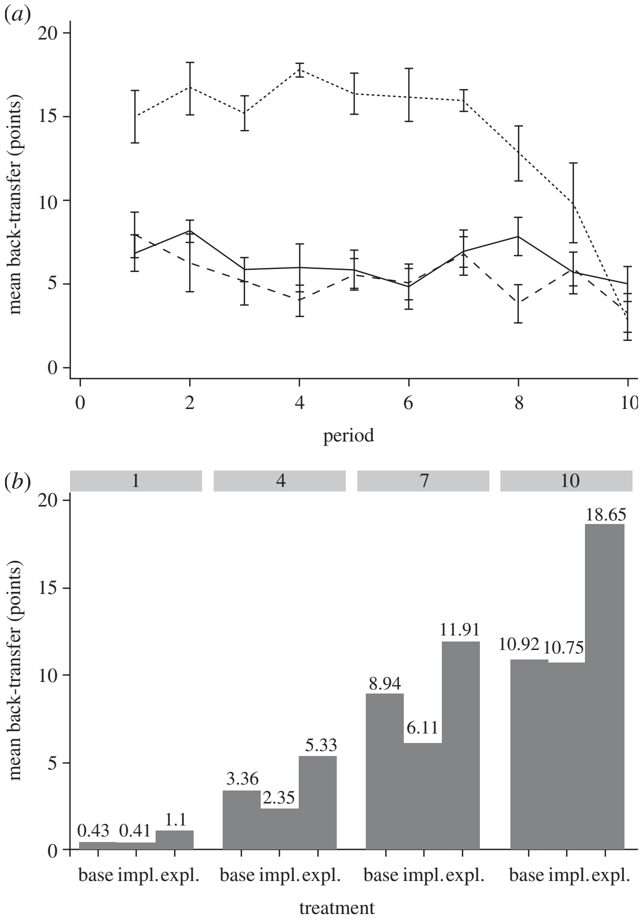


Figure 2. Trustees' mean back-transfers. (*a*) Over time across treatment conditions; error bars represent s.e. on matching group level ($n = 24$). Solid line, baseline; dashed line, eyespots; dotted line, explicit. (*b*) Per investment level across treatment conditions.

Figure 2*a* shows the time path of average back-transfers. The figure indicates that the average back-transfer varies between 5 and 8 units in both the baseline condition and the implicit cue condition, with little difference between the conditions. The figure also displays the s.e.'s (clustered on matching groups) of the mean, which indicate that the differences between baseline and implicit cue conditions are not significant (Mann–Whitney test: $p \geq 0.272$, Holm-Bonferroni correction for multiple comparisons).

Thus, table 2 and figure 2*a* provide little indication that the implicit cue condition increased average back-transfers. An examination of the impact of implicit cues in more detail requires further control for the investments that the trustees face. Figure 2*b* shows the trustees' average back-transfer conditional on the received investments. On an average, trustees in the implicit cues condition sent back the same or a slightly smaller amount than in the baseline condition for any given investment level. In table 3 we report the results of ordinary least squares regressions with the average relative back-transfer as the dependent variable. The relative back-transfer is defined as the share of points returned over the points received (i.e. the quadrupled investment). Thus, a choice of 'nothing' implies a relative back-transfer of 0 per cent, the choice of 'compensate' translates into a relative back-transfer of 25 per cent and the choice of 'equalize' means that 62.5 per cent of the received points are sent back by the trustee. For example, with an investment level of 4, the trustee receives $4 \times 4 = 16$ points and sends back 10 if he chooses 'equalize', giving a payoff of 16 to each of the two players; the relative back-transfer equals $10/16 = 62.5$ per cent. Likewise, if the trustor sends 10, the trustee receives 40 and sends back 25 in the case of 'equalize', which yields a relative back-transfer of $25/40 = 62.5$ per cent. The advantage of using relative back-transfers is that a given choice, such as 'equalize', implies the same percentage number regardless of the investment level. Thus, our regressions implicitly estimate the conditional frequency of the three choices—'nothing', 'compensate' and 'equalized'.

Model (1) in table 3 reports the result of a regression that takes the average relative back-transfer per matching group as the dependent variable. The independent variables in this regression are the average investment per matching group, dummy variables for the implicit cue and the explicit reputation treatment, the average Mach score of the trustees in the matching group, and the average response to the fairness question (high answer indicates low fairness norm). In all regressions, the omitted category is the baseline dummy, implying that the constant measures the average relative back-transfer in the baseline condition, while the dummy for the implicit cue (explicit reputation) condition measures the difference between the baseline condition and the implicit cue (explicit reputation) condition.

Regression (1) is the most conservative because the unit of observation is average behaviour in a matching group, giving us 'only' 24 observations in total. We find a highly significant positive effect of the investment level, i.e. higher investments generate higher relative back-transfers. For our purposes, the most important result of regression (1) is the small and insignificant effect of the dummy for the implicit treatment. The coefficient for this dummy is close to zero, highly insignificant ($p = 0.533$) and even has the 'wrong' sign, indicating that eyespots certainly have no positive effect on trustees' back-transfers. In addition, we find a significant ($p = 0.021$) effect of the fairness standard in the baseline condition—subjects with a lower fairness standard tend to make lower back-transfers.

In regression (2), we examine the mean relative back-transfer on the individual level. This yields 144 observations, as we have 36 trustees in the explicit reputation treatment ('explicit') and 54 in each of the other two conditions (s.e.s are clustered at the matching

Table 3. OLS Regression analysis of trustee decisions. *p*-values are given in parentheses; in models (2) and (3), we use Eicker–Huber–White sandwich estimators for the s.e., clustering on matching groups. Models (1) and (2) use aggregated data on the matching group and trustee level, respectively (i.e. the variables' relative back-transfer and 'investment level' are averages on the matching group and trustee level, respectively. 'Mach' and 'fairness' represent matching group averages in model (1). Regressors: 'investment level' is the number of points the trustor transfers; 'implicit' and 'explicit' are dummies for the respective treatments (omitted category: baseline treatment). In models (2) and (3), 'Mach' is a dummy variable that equals 1 if the participant scored above the median in the Mach-IV inventory. In models (2) and (3), 'fairness' is a dummy variable that equals 1 if the participant's answer to the fairness question was above the median response (i.e. if the subject's fairness standard is below the median). 'explicit × last 3' is a dummy that equals 1 if the observation comes from periods 8, 9 or 10 in the explicit treatment. Its purpose is to capture the end-game effect that occurs when the future benefits from reputation vanish towards the end of the experiment. 'Period' denotes the experimental period and ranges from 1 to 10.

dependent variable	(1) mean relative back-transfer at the level of the matching group		(2) mean relative back-transfer at the level of the trustee		(3) relative back-transfer in individual decisions	
(mean of) investment level	0.031	(0.001)	0.033	(0.000)	0.018	(0.000)
implicit	−0.013	(0.533)	−0.040	(0.242)	−0.038	(0.304)
explicit	0.170	(0.000)	0.107	(0.012)	0.203	(0.000)
(mean of) Mach score	0.062	(0.131)	−0.086	(0.056)	−0.081	(0.076)
Mach × implicit			0.061	(0.279)	0.057	(0.319)
Mach × explicit			0.110	(0.033)	0.105	(0.044)
(mean of) fairness	−0.108	(0.021)	−0.112	(0.056)	−0.109	(0.054)
fairness × implicit			−0.000	(0.998)	−0.006	(0.921)
fairness × explicit			0.007	(0.919)	−0.018	(0.774)
explicit × last 3					−0.176	(0.000)
period					−0.006	(0.001)
constant	0.049	(0.347)	0.113	(0.019)	0.230	(0.000)
n/no. of variance clusters	24/—		144/24		1440/24	

group level). In models (2) and (3), Mach score and Fairness represent dummies for subjects with an above-median Mach score and a below-median fairness standard; recall that subjects with a below-median fairness standard are those who give a high (above-median) rating on the fairness question. Interestingly, this has little effect on the impact of the investment level and implicit cue condition: we get virtually the same result as in model (1), with respect to both the size and the significance of these coefficients. In particular, the coefficient of the implicit cue treatment is still very small, insignificant ($p = 0.242$) and has the wrong sign. However, because of the larger number of observations, the fairness standard and individuals' Mach score is now almost significant at the 5 per cent level (both $p = 0.056$); subjects with lower fairness standards and a higher Mach score back-transfer less in the baseline condition. We are also able to examine the interaction between the fairness standard, the Mach score and the implicit cue condition in regression (2). The interaction between the fairness standard and the implicit cue condition is clearly insignificant ($p = 0.998$); the same holds for the Mach score ($p = 0.279$). This indicates that the implicit cue condition also does not cause behavioural changes in trustees with different fairness standards and different Mach scores.

Finally, we take the decisions in each period as units of observation and cluster again on matching groups in model (3). The dependent variable is now the individual relative back-transfer in a period, which limits the observations to 0, 25 or 62.5 per cent of the received amount. In model (3), we also include variables that capture time effects.

The results of model (3) are interesting in several respects. First, and most importantly, the coefficient for the implicit cue treatment remains small in magnitude and insignificant ($p = 0.304$), and again has the wrong sign. Second, subjects who have a low fairness standard contribute less in the baseline condition ($p = 0.054$). Third, subjects with a high Mach score also contribute less in the baseline condition ($p = 0.076$). Both the second and the third effect are substantial, reducing the mean relative back-transfer by between 8 and 11 per cent. Fourth, the interaction between the implicit cue condition and the below-median fairness standard/above-median Mach dummy is not significant ($p = 0.921/p = 0.319$), indicating that individuals with a low fairness standard/high Mach score do not respond differently to the implicit reputation cue compared with individuals with a high fairness standard/low Mach score. Thus, there is no evidence that individuals who score low on selfishness and opportunism are more prone to implicit reputation cues. Both high and low Mach individuals show little response to the implicit reputation cue.

5. THE IMPACT OF EXPLICIT REPUTATION

In this section, we examine the effect of pecuniary reputation incentives on the trustees' back-transfers and the trustors' investments. Table 2 shows that—in contrast to the implicit reputation condition—the explicit reputation condition causes an enormous increase in average back-transfers—from 6.28 to 13.86. While the modal response in the baseline condition is 'nothing', the modal response in the explicit reputation condition is 'equalize'. This big change in the trustees' back-transfers is highly significant (Mann–Whitney test: $p = 0.002$) and led to a significant increase in the trustors' investments—from 5.88 to 7.74 (Mann–Whitney test: $p = 0.006$). In the explicit condition, the maximum investment also represents the median investment choice. This strong impact of pecuniary reputation incentives can also be seen in figure 2*a*. The average back-transfer is much higher in the explicit reputation condition in all but the last few periods. The time path of the average back-transfer in figure 2*a* also indicates the relatively high degree of rationality that seems to be present in our experiment. During the early periods, a high back-transfer generates a good reputation for many remaining periods, implying that the pecuniary return of a good reputation is high. During the final few periods, a high back-transfer generates a good reputation only for a few remaining periods, implying that the pecuniary return of a good reputation is lower. Thus, individuals who understand this should choose lower back-transfers during the final few periods because the selfish returns of behaving in this way are lower. The time pattern of back-transfers in figure 2*a* is consistent with this rational choice argument.

Interestingly, in period 10 of the explicit reputation condition—in which there are no pecuniary reputation incentives at all—the average level of back-transfers is very similar to the level in the other two conditions (in which explicit reputation incentives are absent by design). Thus, the trustees seem to understand the logic of pecuniary reputation incentives quite well: while they do not respond to merely implicit reputation cues that carry no explicit incentive power, they respond strongly to explicit reputation incentives, and they seem to understand when they can gain from a good reputation and when not.

The powerful effect of pecuniary reputation incentives can also be seen in figure 2*b*, which controls for trustors' investments: trustees' back-transfers are higher at every investment level than in the other two conditions.

Finally, the regressions in table 3 provide further statistical support for the large effect of the explicit reputation condition. In models (1) and (2), the explicit reputation incentive increases the average relative back-transfer by 17.6 and 16.2 per cent, respectively ($p < 0.001$ and $p = 0.012$). Note that in model (3) the inclusion of the 'explicit × last 3' interaction implies that the 'explicit' variable captures the effect of the explicit reputation incentive for the first seven periods while the variable 'explicit × last 3' measures the decrease of back-transfers during the final three periods. The coefficient of 0.203 ($p < 0.001$) for the variable 'explicit' thus indicates that in the first seven periods subjects increase the relative back-transfer relative to the baseline condition by 20.3 per cent if they face an explicit reputation incentive. Moreover, highly opportunistic trustees (above-median Mach score) show a 10.5 per cent larger increase in relative back-transfers when they face an explicit incentive (coefficient of 0.105; $p = 0.044$). Taken together, these results indicate a large effect of the explicit reputation incentive—an effect that contrasts sharply with the null effect of the implicit reputation cue. In fact, an *F*-test indicates that the difference between the coefficients of the implicit and the explicit condition is highly significant ($p < 0.001$).

The above results confirm the hypothesis that Machiavellian subjects respond particularly strongly to social punishment threats such as loss of reputation. This finding is consistent with the results of another study (Simpson & Willer 2008), which also observes that egoistic subjects show a stronger response to pecuniary reputation incentives.

6. DISCUSSION

There is little disagreement among researchers that explicit reputation incentives strongly affect human prosocial behaviour. These explicit incentives can take the form of higher future material benefits in a dynamic experimental game—such as in our explicit reputation treatment—or they can arise when real people (e.g. an audience) saliently observe other people's cooperative or non-cooperative behaviour (Gächter & Fehr 1999; Rege & Telle 2004; Kurzban *et al.* 2007; Smith *et al.* 2009). The strength of merely implicit reputation cues, in which subjects cannot really acquire a good or bad reputation, is, however, much less investigated.

Therefore, we examined the impact of such cues on the strongly reciprocal behaviour of trustees in a trust game. Previous work has argued that eye cues activate reputational concerns, but did not explicitly compare the effect of eye cues with the effect of explicit pecuniary reputation incentives. If reputational concerns shaped humans' altruistic inclinations in ancestral environments to the extent suggested in some of the recent literature (Haley & Fessler 2005; Burnham & Hare 2007)—that is, if humans are indeed so sensitive to reputation cues that they respond to them even if they carry, in fact, no real pecuniary incentive power—subjects should generate patterns in the eye cue condition that resemble the effects of explicit pecuniary reputation incentives.

However, our results indicate that eye cues, which have been hypothesized to represent reliable indicators of potential observability of one's behaviour over the course of human evolution, have no effect at all on the trustees' altruistic behaviour. The effect of the implicit cues treatment is close to zero, highly insignificant and even has the wrong sign. Moreover, this null effect holds regardless of whether we examine the response of subjects who score high or low on the Mach scale. Our results therefore suggest an extremely cautious view of claims that most of the observed prosocial behaviour in anonymous one-shot games should be attributed to uncontrolled implicit reputation cues. At the current state of our knowledge, this claim represents no more than a speculation, lacking empirical support. If it were indeed the case that uncontrolled reputation cues are so important, behaviour should also respond to experimentally controlled implicit reputation cues.

The null effect of the implicit reputation cue contrasts sharply with the large impact of explicit pecuniary reputation incentives on trustees' behaviour. The large contrast between the implicit and the explicit reputation condition reinforces our conclusions above. The effect of the implicit cue does not even resemble the effect generated by the pecuniary reputation incentive, suggesting that implicit cues are a relatively weak force.

We also found important individual differences in subjects' responses to the pecuniary reputation incentive.

Subjects who score high on the Mach scale behave less altruistically in the baseline treatment, but they respond more strongly to the pecuniary reputation incentive.

Why do other studies find an effect of eye cues on prosocial behaviour while we find none? With regard to the studies of Bateson *et al.* (2006) and Burnham & Hare (2007), the following feature of their experiments might have caused the difference. Both experiments investigate contributions to a public good. As shown above, many people are conditional cooperators, and their contributions therefore depend on their beliefs about other people's contributions. Eye cues could generate more optimistic beliefs about other subjects' cooperation behaviour, which then induce higher cooperation rates among subjects with preferences for conditional cooperation. This contrasts with our study in which we have full control over subjects' beliefs because the trustees know exactly the investment level if they make their back-transfer. Therefore, in our study, eye cues cannot affect beliefs about other subjects' behaviour.

With regard to the study of Bateson *et al.* (2006)—a field experiment about voluntary contributions to an honesty box in a university coffee room—another feature is also potentially important. Subjects often consume coffee jointly and observe whether their colleagues pay for the coffee. In this case the subject's real reputation—and not just its imagined reputation—is at stake. If eye cues draw attention to the moral appropriateness of paying for one's coffee, then this real reputation incentive may be greatly strengthened. Thus, it is possible that the eye cues in the Bateson *et al.* experiment enhanced the already prevailing incentive to maintain one's reputation as an honest coffee consumer. This feature of the Bateson *et al.* experiment also contrasts with our experiment because we rule out any interaction between the eye cue and the pecuniary (explicit) reputation incentive.

Why did the eye cue affect the behaviour in the dictator game experiment of Haley & Fessler (2005) while lacking effect in our trust game? A possible reason for this may be that the dictator game constitutes a less robust situation. Experimental economists now generally acknowledge that the dictator game is likely to involve more experimenter demand effects (Bardsley 2008) and is less robust than other games in which subjects interact with each other (Cooper & Kagel 2010). Therefore, relatively weak forces can affect behaviour in the dictator game. Perhaps the implicit reputation cue is one of these weak forces.

The authors would like to thank Charles Efferson, Ryan McKay, Sonja Vogt and two excellent referees for their many useful comments on the manuscript.

REFERENCES

Bardsley, N. 2008 Dictator game giving: altruism or artefact? *Exp. Econ.* **11**, 122–133. (doi:10.1007/s10683-007-9172-2)

Bateson, M., Nettle, D. & Roberts, G. 2006 Cues of being watched enhance co-operation in a real-world setting. *Biol. Lett.* **2**, 412–414. (doi:10.1098/rsbl.2006.0509)

Bowles, S. & Gintis, H. 2004 The evolution of strong reciprocity: co-operation in heterogeneous populations. *Theoret. Popul. Biol.* **65**, 17–28. (doi:10.1016/j.tpb.2003.07.001)

Boyd, R. T. & Richerson, P. 2005 *The origin and evolution of cultures*. Oxford, UK: Oxford University Press.

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. 2003 The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535. (doi:10.1073/pnas.0630443100)

Brown, M., Falk, A. & Fehr, E. 2004 Relational contracts and the nature of market interactions. *Econometrica* **72**, 747–780. (doi:10.1111/j.1468-0262.2004.00511.x)

Burnham, T. C. & Hare, B. 2007 Engineering human co-operation: does involuntary neural activation increase public goods contributions? *Human Nat.* **18**, 88–108. (doi:10.1007/s12110-007-9012-2)

Camerer, C. F. 2003 *Behavioural game theory—experiments in strategic interaction*. Princeton, NJ: Princeton University Press.

Christie, R. & Geis, F. 1970 *Studies in Machiavellianism*. New York, NY: Academic Press.

Cochard, F., Van, P. N. & Willinger, M. 2004 Trusting behavior in a repeated investment game. *J. Econ. Behav. Org.* **55**, 31–44. (doi:10.1016/j.jebo.2003.07.004)

Cooper, D. & Kagel, J. H. 2010 Other-regarding preferences: a selective survey of experimental results. In *Handbook of experimental economics* (eds A. Roth & J. H. Kagel). Princeton, NJ: Princeton University Press.

Croson, R. 2007 Theories of commitment, altruism and reciprocity: evidence from linear public goods games. *Econ. Inquiry* **45**, 199–216.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R. & Smirnov, O. 2007 Egalitarian motives in humans. *Nature* **446**, 794–796. (doi:10.1038/nature05651)

Dufwenberg, M. & Kirchsteiger, G. 2004 A theory of sequential reciprocity. *Games Econ. Behav.* **47**, 268–298. (doi:10.1016/j.geb.2003.06.003)

Engelmann, D. & Fischbacher, U. 2009 Indirect reciprocity and strategic reputation building in an experimental helping game. *Games Econ. Behav.* **67**, 399–407. (doi:10.1016/j.geb.2008.12.006)

Falk, A. & Fischbacher, U. 2006 A theory of reciprocity. *Games Econ. Behav.* **54**, 293–315. (doi:10.1016/j.geb.2005.03.001)

Fehr, E. & Fischbacher, U. 2003 The nature of human altruism. *Nature* **425**, 785–791. (doi:10.1038/nature02043)

Fehr, E. & Schmidt, K. M. 1999 A theory of fairness, competition, and co-operation. *Q. J. Econ.* **114**, 817–868. (doi:10.1162/003355399556151)

Fehr, E., Fischbacher, U. & Gächter, S. 2002 Strong reciprocity, human co-operation, and the enforcement of social norms. *Human Nat.* **13**, 1–25. (doi:10.1007/s12110-002-1012-7)

Fehr, E., Brown, M. & Zehnder, C. 2009 On reputation: a microfoundation of contract enforcement and price rigidity. *Econ. J.* **119**, 333–353. (doi:10.1111/j.1468-0297.2008.02240.x)

Fischbacher, U., Gächter, S. & Fehr, E. 2001 Are people conditionally co-operative? Evidence from a public goods experiment. *Econ. Lett.* **71**, 397–404. (doi:10.1016/S0165-1765(01)00394-9)

Gächter, S. & Falk, A. 2002 Reputation and reciprocity: consequences for the labour relation. *Scand. J. Econ.* **104**, 1–27. (doi:10.1111/1467-9442.00269)

Gächter, S. & Fehr, E. 1999 Collective action as a social exchange. *J. Econ. Behav. Org.* **39**, 341–369. (doi:10.1016/S0167-2681(99)00045-1)

Gintis, H. 2000 Strong reciprocity and human sociality. *J. Theoret. Biol.* **206**, 169–179. (doi:10.1006/jtbi.2000.2111)

Gintis, H., Bowles, S., Boyd, R. & Fehr, E. 2003 Explaining altruistic behaviour in humans. *Evol. Hum. Behav.* **24**, 153–172. (doi:10.1016/S1090-5138(02)00157-5)

Hagen, E. H. & Hammerstein, P. 2006 Game theory and human evolution: a critique of some recent interpretations of experimental games. *Theoret. Popul. Biol.* **69**, 339–348. (doi:10.1016/j.tpb.2005.09.005)

Haley, K. J. & Fessler, D. M. T. 2005 Nobody's watching? Subtle cues can affect generosity in an anonymous economic game. *Evol. Hum. Behav.* **26**, 245–256. (doi:10.1016/j.evolhumbehav.2005.01.002)

Henrich, J. & Boyd, R. 2001 Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in co-operative dilemmas. *J. Theoret. Biol.* **208**, 79–89. (doi:10.1006/jtbi.2000.2202)

Kocher, M. G., Cherry, T., Kroll, S., Netzer, R. & Sutter, M. 2008 Conditional co-operation on three continents. *Econ. Lett.* **101**, 175–178. (doi:10.1016/j.econlet.2008.07.015)

Kurzban, R. & Houser, D. 2005 Experiments investigating cooperative types in humans: a complement to evolutionary theory and simulations. *Proc. Natl Acad. USA* **102**, 1803–1807. (doi:10.1073/pnas.0408759102)

Kurzban, R., DeScioli, P. & O'Brien, E. 2007 Audience effects on moralistic punishment. *Evol. Hum. Behav.* **28**, 75–84. (doi:10.1016/j.evolhumbehav.2006.06.001)

Milinski, M., Semmann, D., Bakker, T. C. M. & Krambeck, H. J. 2001 Co-operation through indirect reciprocity: image scoring or standing strategy? *Proc. R. Soc. Lond. B* **268**, 2495–2501. (doi:10.1098/rspb.2001.1809)

Milinski, M., Semmann, D. & Krambeck, H. J. 2002 Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426. (doi:10.1038/415424a)

Rabin, M. 1993 Incorporating fairness into game theory and economics. *Am. Econ. Rev.* **83**, 1281–1302.

Rege, M. & Telle, K. 2004 The impact of social approval and framing on co-operation in public good situations. *J. Public Econ.* **88**, 1625–1644. (doi:10.1016/S0047-2727(03)00021-5)

Rigdon, M., Ishii, K., Watabe, M. & Kitayama, S. 2009 Minimal social cues in the dictator game. *J. Econ. Psychol.* **30**, 358–367. (doi:10.1016/j.joep.2009.02.002)

Rockenbach, B. & Milinski, M. 2006 The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444**, 718–723. (doi:10.1038/nature05229)

Simpson, B. & Willer, R. 2008 Altruism and indirect reciprocity: the interaction of person and situation in prosocial behavior. *Soc. Psychol. Quart.* **71**, 37–52.

Smith, F. G., Debruine, L. M., Jones, B. C., Krupp, D. B., Welling, L. L. M. & Conway, C. A. 2009 Attractiveness qualifies the effect of observation on trusting behavior in an economic game. *Evol. Hum. Behav.* **30**, 393–397. (doi:10.1016/j.evolhumbehav.2009.06.003)

Spitzer, M., Fischbacher, U., Herrnberger, B., Gron, G. & Fehr, E. 2007 The neural signature of social norm compliance. *Neuron* **56**, 185–196. (doi:10.1016/j.neuron.2007.09.011)

Wedekind, C. & Milinski, M. 2000 Co-operation through image scoring in humans. *Science* **288**, 850–852. (doi:10.1126/science.288.5467.850)