Preliminary – do not cite without permission

# Implicit Reputation Cues and Strong Reciprocity

An Experimental Study

April 21, 2007

*Ernst Fehr* *Frédéric Schneider*

Institute for Empirical Research in Economics, University of Zürich

### Abstract

Evolutionary psychologists argue that strong reciprocity is primarily shaped by an evolutionary history of repeated interactions in which it was in the self-interest of people to reciprocate favors. For this reason modern humans are hypothesized to respond to variations in all kinds of subtle reputational cues such as human voices in the environment, whether actions take place in the dark or during daylight, i.e, by cues that induce feelings of being observed. These variations in reputational cues are predicted to affect reciprocal behavior even if they are *not* associated with any changes in pecuniary incentives. We address this hypothesis by implementing an implicit reputation treatment in which trustees in a trust game are "observed" by eyes on their computer screen which are thought to activate emotional programs of prosocial behavior. In order to assess the relative importance of potential reciprocity-enhancing effects of implicit reputational cues we compare this treatment with a treatment in which trustees have a pecuniary incentive to behave nicely because the future trustors are informed about the trustees past actions. We find no support for the evolutionary psychology hypothesis because implicit reputation cues have no effect on trustee behavior while explicit pecuniary reputation incentives cause large increase in trustees' back transfers.

## Contents

## List of Figures

## List of Tables

# 1 Introduction

Standard economics has always assumed that the driving force behind human behavior is material payoff maximization. By now, it is a well established fact that this is not always true. Economic experiments in the laboratory and in the field show that humans deviate from purely selfish behavior in different ways, giving up monetary benefits to reward or punish others[1]. This kind of social concerns cannot be explained by purely strategic cooperation, since humans display cooperativeness even in anonymous one-shot interactions in which cooperation (or punishment) does not entail any strategic advantage. Behavioral economics try to account for social concerns via modified preference structures[2]. Social preferences can explain the findings of pro-social behavior in non-strategic situations. In the case of reciprocal behavior, this is called *strong reciprocity*[3], in contrast to weak reciprocity. Weak reciprocity occurs in repeated game situations where mutual cooperation is in line with strategical self-interest.

Some evolutionary psychologists argue that social preferences are an epiphenomenon of strategical interest. While social preferences are a possible proximate explanatory mechanism for strong reciprocity, they propose[4] that the ultimate mechanism that leads to strong reciprocity can be explained by means of kin selection, reciprocal altruism and indirect reciprocity. Humans do not reveal intrinsic social preferences which are evolutionarily plausible by themselves[5], but rather apply social heuristics that were fitness enhancing in the past, but are actually a fitness penalty in the modern context of one-shot interactions between strangers. Helping others was – on an evolutionary time scale – an optimal way of building a "positive image score", i. e., a good reputation, in order to get help in the future. This sort of strategic cooperation is called *indirect reciprocity*. Contrary to weak reciprocity, indirect reciprocity can also emerge in one-shot situations, provided that they are not anonymous, i. e., the deed is publicly observable, thus increasing the

[1] see Fehr and Fischbacher (2003) and Fehr and Gächter (2002a), for an overview on experimental evidence: Roth (1995)

[2] the most prominent are Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), Bolton and Ockenfels (2000), Fehr and Schmidt (1999), Charness and Rabin (2002)

[3] The concept was first outlined by Trivers (1971); for a definition and experimental exploration, see Fehr et al. (2003)

[4] see Heintz (2005), Johnson et al. (2003) for the argument that follows

[5] despite the existence of models that give an evolutional rationale for strong reciprocity on the basis of group selection, see Gintis (2000)

standing in the eyes of others. This does not explain why a considerable fraction of people tend to be "nice" even in *anonymous* one-shot interactions. Therefore, proponents of indirect reciprocity assert that empirically observed strong reciprocity is actually misguided indirect reciprocity. Their rationale is the following: the different functions of the human mind have been shaped by evolution[6]; evolutionary adaptation is a slow process that takes many generations to manifest; large human societies composed of genetically and socially loosely related individuals have emerged only recently on the evolutionary time scale; this means, human social behavior is well adapted to prehistoric tribal societies[7]; social life, it is presumed, took place in small groups of kin individuals; any social activity was more likely to be observed than in modern days; furthermore, the reputational consequences were supposedly important, because the observer was a future interaction partner.

From this (speculative) scenario evolutionary psychologists conclude the following concerning strong reciprocity: In modern days, humans assess the a-priori probability of being watched as well as the arising reputational consequences exceedingly high. As a result, they tend to over-react to cues of being observed, such as bright daylight, or sounds of others being around. Strong reciprocity is in fact only a reaction of our prehistoric brain on cues that do not have any relevance in *today's* social life of one-shot interactions between strangers, but were of crucial importance in our tribal past. In the paper, we will refer to this rationale as the "Maladaption Theory"[8].

This raises the question how to disentangle the two alternative ultimate explanations for strong reciprocity – emergence of social preferences through group selection or maladapted self-interest through evolutionarily shaped social heuristics. It is necessary to find a critical point where both theories make distinct predictions on human behavior. The "implicit cues hypothesis" seems to be promising in this respect: while "intrinsic" social preference models do not only predict behavioral changes when reputational consequences are obvious, maladaption predicts an increase in pro-social behavior even in the presence of so called "subtle" or "implicit" cues, i. e. cues that do not enter the decision-making process via cognition but rather via an affective channel. It has been proposed that the pictorial representation of a face acts

[6] The so called "Modular Mind" Hypothesis was formulated by Fodor (1983)

[7] For an outline of the stance of evolutionary psychology on human social behavior, see Barkow et al. (1992)

[8] see Fehr and Henrich (2003)

as such a cue[9]. It is clearly identifiable as reputationally non-relevant but is presumed to raise the affective judgment of being watched. Experiments involving implicit cues have already been conducted.

Haley and Fessler (2005) administered participants of a dictator game experiment either a neutral control wallpaper on the computer screen background or a face like shape with stylized eyespots. They report a significantly higher proportion of dictators who transfer a positive amount when they faced the eyespots. They conclude that prosocial behavior in anonymous one-shot games can be explained by implicit "reputational factors" that are uncontrolledly present during experiments.

Bateson et al. (2006) conducted a field experiment on contributions to a public good. In the coffee break room of the psychology department at the University of Newcastle, they printed an image on the notice that reminded the users to put money into the honesty box according to the amount of coffee they consume. Each week, the image alternated between a control (flowers) and different pairs of eyes – which could be of either sex and have a different expression, ranging from seductive to angry. Most of the treatment images resulted in higher contributions than the control images.

This second experiment shows nicely the practical interest for behavioral economics. There is an abundance of situations with imperfect information in economic life, for example in work relationships where agents' actions are unverifiable for principals, when purchasing consumer goods where quality is difficult to assess by the buyer[10], or in financial markets where lenders cannot monitor borrower behavior[11]. Thus, economic relationships where the uninformed party can trust the informed party are of great benefit because trust can potentially alleviate the inefficiencies arising from imperfect information[12]. One way to impose good conduct[13] in *repeated* one-shot situations even on selfish agents is the institution of reputation in the form of information on the agents' previous behavior. If there is a way to effectively

[9] Haley and Fessler (2005)

[10] in online auctions: see Livingston (2005) for a model and Resnick et al. (2006) for evidence from a field experiment.

[11] Brown and Zehnder (2005) examine the relevance of borrower reputation for repayment rates and market performancein a laboratory experiment

[12] for experimental results on how trustful environments can positively affect principal-agent relationships see Falk and Kosfeld (2006)

[13] In the form of rational cooperation; in repeated games, the sequential equilibrium strategy of selfish agents is cooperation, see Kreps et al. (1982) and Andreoni and Miller (1993) for an in depth discussion.

increase prosocial behavior by applying implicit cues in the framework of work relationships, this will have considerable impact on company policies.

The experiments conducted by Haley/Fessler and Bateson et al. are not conclusive in answering the question for the ultimate mechanism of strong reciprocity for several reasons. First of all, Haley/Fessler conducted a dictator game. This game is not suitable to examine reciprocity because there is just one player – the dictator – with an action space while the other has only a passive role. This means that the dictator does not reciprocate anything; rather, giving in the dictator game is an act of unconditional altruism. Moreover, the dictator game is sensitive to framing issues, precisely because the context is not self-explanatory. The dictator does only know that she is endowed with points/money from the experimenter which makes the game susceptible to experimenter demand effects. Second, the game was played only once. It may be that the presumed social heuristic which is unconsciously applied can be overcome by learning. Bateson et al. conduct a field experiment in a public goods setting. Here reciprocity is in principle possible since one's own contribution reciprocates the others' contributions. However, the participants did not know anything about the contributions of the others – neither the overall size, nor the distribution of individual contributions. Therefore, they had to rely on their beliefs which may have even been updated somehow in the presence of the different visual cues. Without controlling for the participants' beliefs, one can not determine the reputational effect of the cues applied.

Furthermore, both experiments compare the effect of implicit cues to a baseline condition without cues. Since these cues are presumed to have an effect on the judgement of the reputational relevance of the interaction, the question remains how big the effect is compared to a situation where participants can build a *real* reputation. Therefore, the relative importance of the effect remains unclear.

The experiment reported in this article was conducted to examine the robustness and importance of the implicit cues effect. Therefore, we conducted a series of one-shot investment games (trust games) which comprises an initial transfer from the so called trustor to the trustee, and in a second stage, a back-transfer from the trustee to the trustor. The trust game has the advantage that it really deals with reciprocity because the trustee reacts to the trustor's action. The situation is clearer than in the dictator game since the trustee receives the money from the trustor and thus, the act of giving money to the other player is less open to interpretation. The trustee knows the size of the trustor's transfer, hence, there is no issue of belief formation as in the Bateson et al. case. In addition, the repetition of the game allows us to observe if trustees diminish their transfers over time.

It is important to emphasize that the existence of genuine social preferences beyond the scope of immediate kinship should not be mistaken for an absence of strategical altruism to build good reputation and status[14]. Doubtlessly, prosociality emerges out of multiple motives. We challenge the view that all instances of prosocial behavior are driven by purely strategical motives, may they be rational or irrational.

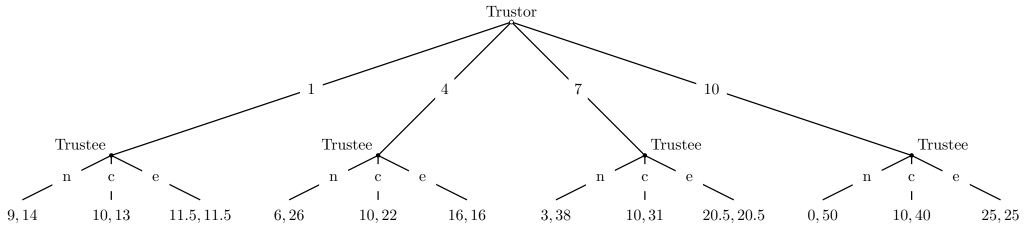
# 2 Experiment

## 2.1 Experimental procedure

We measured strong reciprocity as second mover behavior in a series of one-shot trust games. The experimental design includes three treatments: a *baseline* treatment where the trustee faces a neutral background screen; an "implicit cues" treatment where the background screen features eyespot-like shapes (we will refer to this treatment as the *eyesspots* treatment or "eyes" for short); and an "explicit reputation" treatment where the trustee's previous decisions are observable to the current trustor ("explicit"). Over all, we conducted 8 sessions (3 sessions in baseline and implicit, 2 in explicit treatment), each session involving 36 participants, 288 participants in total (144 trustors and 144 trustees). In their instructions, subjects were given the role of a trustor or a trustee. Then, they played 10 periods of the trust game, with randomly re-matched partners each period. Immediately after the end of the last period, the participants had to fill out a questionnaire on different issues (emotional state, fairness attitudes, Machiavelli, trust, socioeconomic data). After completion, participants were paid a show-up fee of 10 Swiss Francs plus the amount of points they earned during the experiment times .2 Swiss Francs.

[14] For experimental evidence of this kind of altruistic behavior, see Hardy and Van Vugt (2006)

Fig. 1: Extensive form of the trust game



## 2.2 Game Design

Each period of the experiment was a one-shot trust game. At the beginning of the period, all participants, 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 wanted to transfer to the trustee with whom they were currently matched; the invested points were then quadrupled and transferred to the trustee; and a back-transfer stage where trustees had to decide how much they wanted to transfer back to the trustor. The strategy spaces were discrete: trustors could choose between 4 possible transfers: 1 point, 4 points, 7 points or 10 points; trustees had three options: they could transfer back nothing, the amount sent by the trustor (1, 4, 7 or 10 points), or they could transfer back an amount that equalized the period payoff between trustor and trustee[15]. Note that the trustee is perfectly informed about the trustor's choice and thus, does not have to form beliefs about the size of the gift he or she might want to reciprocate. This information is not given in the field experiment of Bateson et al. (2006), and there is no control for the participant's beliefs.

The reason why we did not include a "zero investment" choice was because we wanted the trustors to send a positive amount in order to render the back-transfer options meaningful (in the case of no investment, every trustee choice results in a zero back-transfer); this was important for the explicit treatment because an investment of zero would have been an occasion for the trustees to build reputation costlessly by choosing "equal split".

Figure 1 shows the corresponding game tree, table 1 the payoff matrix.

[15] Henceforth, we will refer to these options as "nothing", "compensate" and "equal split"/"equalize". Not that these terms were *not* used in the experiment.

Tab. 1: Payoff matrix of the trust game

| | | Trustee | | |
|---|---|---|---|---|
| | | nothing | compensate | equalize |
| Trustor | 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 |

## 2.3 Predictions

Consider the two anonymous treatment conditions first. There the individual periods are true one-shot games. The unique Nash equilibrium is minimal investment and zero back-transfer. A rational selfish trustee should never transfer a positive amount since these strategies are strictly dominated. A rational trustor takes this into account and invests the minimal amount.

There can be two reasons why a trustor may return a positive amount to the trustor. First, the trustee can have preferences that make a positive transfer utility-maximizing, e. g. inequity aversion. In this case there would be no difference between the baseline treatment and the eyespots treatment. The social preference can be conditional or unconditional on the trustor's decision. While a strongly inequity-averse agent may always choose the "equal split" option, regardless of the received investment, a trustor with a taste for intentional fairness would punish a distrusting first mover with a zero return and reward a trusting first mover by equalizing payoffs. Second, the trustee may suffer from evolutionary maladaptation and assume reputational consequences in a social interaction that is in fact anonymous. In this case, the perceived reputational relevance would be increased in the eyespots treatment and result in higher back-transfers compared to the baseline condition. Note that the subtle cues hypothesis is the only one that predicts this difference.

The third treatment removes anonymity in the respect that trustees have an observable history of back-transfers which can serve the current trustor to make an informed investment decision[16]. This information puts the transaction partners in a situation that resembles more a repeated than a one-shot

[16] for a discussion and experimental analysis of the interaction between reciprocity and reputation, see Gächter and Falk (2002)

game. Selfish trustees should now choose positive back-transfers because rational trustors will invest high amounts in cooperative trustees; but only the minimal amount in non-cooperative trustors [17] – except for the last period. Since there are no future periods which justify the maintenance of the trustee's reputation, the selfish trustor will defect. Rational trustors will anticipate this behavior and restrict their investment to the minimal amount.

## 2.4 Questionnaire

We administered a battery of questionnaires at the end of the experiment, examining momentary emotional state, fairness attitudes, machiavellistic personality, trust attitudes and socio-economic information.

**Emotional State** We included this questionnaire to control for the potential confound that participants may behave differently in the experiment because the presented cues altered their emotional state. There are 19 items of the form: "At the moment, I am ..." where the blank is replaced by an emotional adjective, for example "happy", "anxious", "callous", "angry" and so forth. The answer had to be given on an nine point scale ranging from "not at all" to "very".

**Fairness** The notion of fair behavior in a certain situation may vary between subjects. It would then be desirable to have a measure of individual fairness norms, especially in the game presented in the experiment. We thus proposed three different hypothetical game constellations and asked participants to judge the fairness of the trustee's behavior on a seven point scale from "very unfair" to "very fair". The three constellations were:

- trustor invests 4 points, trustee chooses equal split; we hypothesized that a maximal return on a low investment would be perceived as very fair.
- trustor invests 7 points, trustee returns nothing; this should be thought of as unfair, because a high investment is not rewarded.

[17] remember that the trustee's payoff is 25 points if the trustor invests 10 points and the trustee chooses "equal split", which is higher than any payoff if the trustor invests only 1 point.

- trustor invests 10 points, trustee chooses “compensate”; we thought that this would be somewhat ambiguous to judge for the participants since the adequacy of the “compensation” choice is debatable. We expected a high variation in responses.

Trustees who assessed the hypothetical trustee’s behavior in the third situation as rather fair were supposed to generally transfer back less, and vice versa. The answer on the seven point scale served as a measure for attitude to fairness in the regression analysis.

**Machiavelli** We surveyed participants concerning their tendency to machiavellianism, i. e. the disposition to instrumentalize others to serve their ends, using the MACH-IV psychometric test[18] with a seven point scale. The obtained machiavelli score ranges from $-3$ to $+3$, higher scores indicate stronger machiavellian personality. In the non-reputational treatments, trustees with high scores were supposed to generally return less than those with low scores since they do not care about fairness norms. In the explicit treatment however, they should respond rationally to the reputational incentives, that is, selfishly cooperate at the beginning and defect at the end of the experiment.

**Trust** To assess personal attitudes to the trustworthiness of others and oneself, we included items used in the GSOEP.

**Socio-economic variables** Our experiments routinely comprise a questionnaire on a host of socio-economic data such as gender, age, income, nationality and the like. These factors may have some effect on the behavior in the trust game.

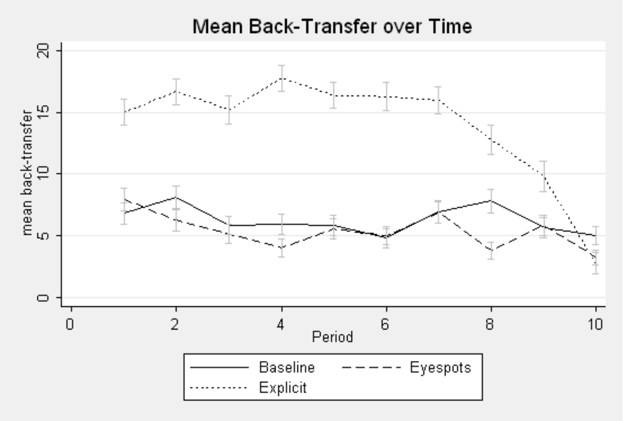
## 3 Results

If one assumes that subtle cues raise reputational concerns, this means that subject behavior should differ from a treatment without exposition to cues, everything else kept constant. Moreover, behavioral patterns should resemble those which occur in a treatment where subjects have the possibility to build a true reputation. Figure 2 gives a first impression of the main results. It shows the average back-transfer over time and treatments. As can easily be

[18] described in Christie and Geis (1970)

seen, there is a big difference between the explicit reputation treatment and the other two treatments, back-transfers being almost double as high in the former condition. Also, a distinct drop at the end of the experiment can be spotted, indicating the presence of an end game effect – but only in the explicit condition. To make this point more obvious, see figure 3, a scatter plot that shows back-transfers (as percentage of points received from the trustor). Each dot represents one trustee; the horizontal axis gives the mean back-transfer over the first seven periods of the experiment, the vertical axis gives the mean back-transfer over the last three periods. The end game effect can be seen in the explicit treatment where almost all dots are below the 45° line, meaning lower back-transfers at the end of the experiment than at the beginning. No such effect is visible in the other two conditions.

Fig. 2: Second mover behavior (error bars: +/- sem)



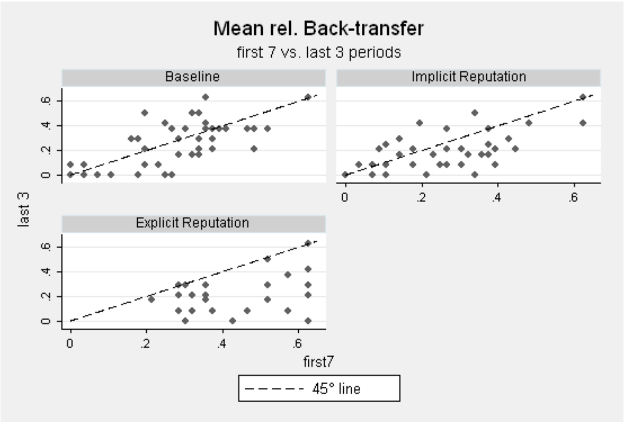
In this section, we establish the behavioral difference between the baseline and the reputation treatment. We examine how the eyespots treatment fits into the picture, both with non-parametric and parametric measures making different assumptions about the way subtle cues may work. Table 2 provides a numeric overview on the experiment.

## 3.1 Questionnaire Data

## 3.2 Non-parametric Results

**Exclusion of Emotion Induction through Visual Stimuli** In our experiment, subjects were exposed to different visual stimuli which can, in principle,

Fig. 3: End game defection



Tab. 2: Summary Statistics

| Statistic | Baseline | Implicit | Explicit |
|---|---|---|---|
| No. subjects | 108 | 108 | 72 |
| No. obs. A | 540 | 540 | 360 |
| No. obs. B | 540 | 540 | 360 |
| No. matching groups | 9 | 9 | 6 |
| Av. points | 188.250 pts. | 186.167 pts. | 215.875 pts. |
| Av. points trustees | 272.491 pts. | 276.167 pts. | 270.417 pts. |
| Av. points trustors | 104.009 pts. | 96.167 pts. | 161.333 |
| Av. Investment | 5.88 pts. | 5.74 pts. | 7.73 pts. |
| Median Investment | 7 pts. | 7 pts. | 10 pts. |
| Av. Back-transfer | 6.28 pts. | 5.36 pts. | 13.86 pts. |
| Av. rel. Back-transf. | 22.66% | 18.89% | 42.08% |
| Mode trustee decision | nothing | nothing | equalize |

induce different emotions. This could potentially flaw our results since effect of the implicit reputation treatment would be indistinguishable from the the emotion induction effect. To exclude this, we administered a questionnaire inquiring the subjects' emotional state immediately after the treatment. The analysis of the questionnaire data shows no influence of the visual cues on the emotional state: for none of the 19 queried emotions, neither Mann-U equality of mean tests nor Kolmogorov-Smirnov tests of equality of distributions reported differences between subjects with and without eyespots on any conventional level, even uncorrected for multiple hypotheses. The lowest p-value for the Kolmogorov-Smirnov tests is $p = 0.413$ ("At the moment, I am upset")[19]. We can therefore exclude an emotion based effect of eyespots on the subjects' behavior.

**Distributions of trustee decisions** We are interested in the differences that may occur by adding subtle reputational cues to a game environment that includes the possibility of strong reciprocity. Of primary interest are, of course, the decisions taken by the trustees. Figure 4 shows the average amount of points trustees sent back conditional on the received investment. The increase with increasing investment is largely an artefact due to the multiplication factor in the payoff equation. Comparisons across investments are therefore difficult. The interesting fact here is that – for every investment level taken separately – the average back-transfer is highest in the reputational treatment while the other two treatments are close together in the extreme investments; and in the intermediate investments, trustees facing eyespots give even less than baseline trustees. However, this phenomenon is not significant on the subject level (Mann-Whitney U on difference in means between baseline and implicit condition gives a p-value of .18) and completely disappears in the regression analysis (cf. section 3.3).

**Distributions of trustor decisions** A more sophisticated issue is a possible difference in *trustor* behavior. Figure 5 shows the distribution of trustor decisions across treatments. Trustors in the two non-explicit treatments had the same information on the game, so we would expect initial behavior to be equally distributed. But according to the subtle cues hypothesis, trustees in the eyespots treatment are supposed to be more generous. And since trustors could update their beliefs on trustee behavior during the experiment, they could have reacted to this generosity by increasing their investments. Yet, since there is no significant difference in trustee behavior, we shouldn't expect a difference in trustee behavior when looking at the non-explicit treatments.

In the explicit treatment, trustors knew that trustees would be disciplined by reputational concerns and invested more (as table 2 shows, the median investment shifted from 7 to 10 points). Moreover, on average trustors correctly anticipated the vanishing of the reputational effect at the end of the

[19] cf. table 5 for a complete list

Fig. 4: Trustee decisions

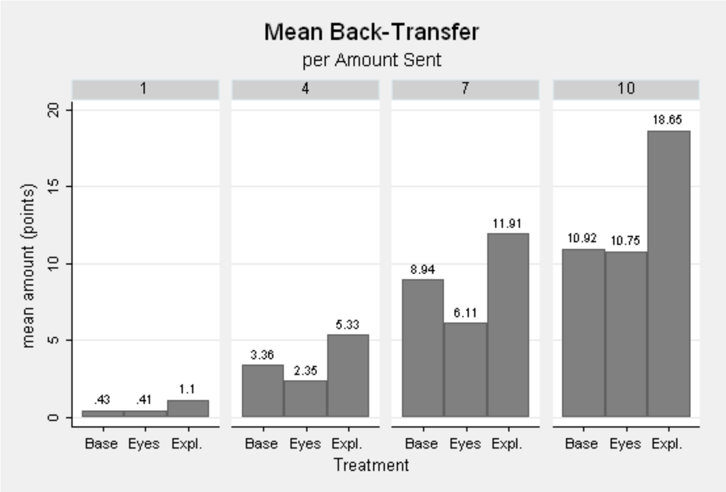
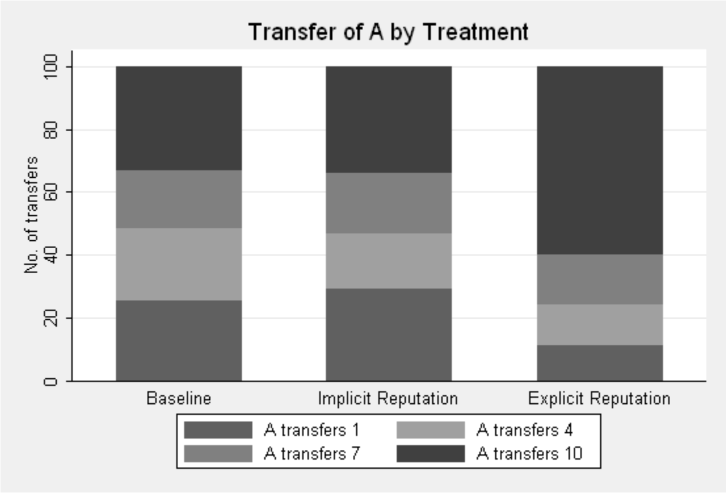
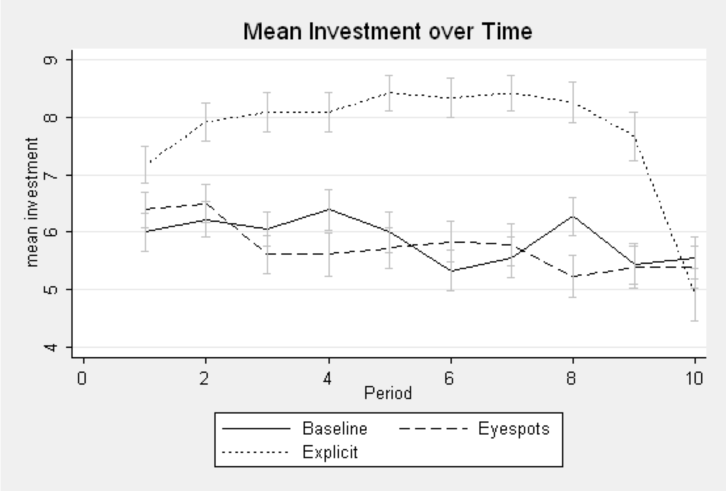


Fig. 5: Trustor decisions



game and reduced investments from 10 points median investment in the second to last period to 4 points median investment in the last period, as can be seen in figure 6.

Fig. 6: Investments over time (error bars = +/- sem)



**Distributions of payoffs** Since the game is deterministic, payoffs are completely determined by the participants' decisions which we examined in the previous paragraphs. Hence, we should see no difference in the distribution of payoffs between Baseline and Implicit treatment, but higher payoffs for trustors in the Explicit treatment because here, trustees transferred much higher amounts back to the trustors. In our setting, reciprocal behavior of trustees leads to efficient payoffs because it induces higher investments on the trustors' side. Thus, the effect of reciprocity on payoffs should be visible in the explicit reputation treatment because here it is rational to act reciprocally. According to the competing hypotheses on strong reciprocity, the behavior in the baseline and implicit reciprocity treatments should either be driven by the same level of *strong reciprocity* or by different levels of (misguided) weak reciprocity. In the latter case, trustees' decisions should deviate from baseline behavior towards behavior under the explicit reputation regime. Looking at the overall payoff distributions in the different treatments, we see a dramatic upward shift from baseline to explicit reputation. On the other hand, the implicit reputation treatment does not change payoffs compared to baseline. If we look at the payoff distributions of trustors and trustees separately, we gain some more insight. Table 3 shows the means and standard errors of payoffs by type and treatment.

Kolmogorov-Smirnov tests show that there is no significant distributional

Tab. 3: Distribution of payoffs by player type

| | Trustor | | Trustee | |
|---|---|---|---|---|
| | mean | s. e. m. | mean | s. e. m. |
| Baseline | 103.9583 | 5.036745 | 269.2917 | 10.12133 |
| Eyespots | 94.3472 | 4.030283 | 277.9028 | 10.96484 |
| Explicit | 161.3333 | 3.866061 | 270.4167 | 5.37651 |
| Total | 119.8796 | 3.786904 | 272.537 | 5.25042 |

difference between baseline and eyespots condition (in fact, trustees earn on average a bit more and trustors a bit less in the eyespots treatment), but a highly significant difference in trustor earnings between eyespots and explicit condition. Trustors in the explicit condition earn about 60% more than trustees in the other treatments. Interestingly, the standard error for trustee payoffs is only half as big in the explicit treatment as in the other two, meaning that there is much less variation in the trustee's earnings because reputational incentives force them to play a cooperative strategy. Also, the mean payoff of trustees does not significantly (Mann-Whitney U, $p = 0.2550$) from the non-explicit treatments. While enjoying a higher level of investments in the explicit condition, trustees are not able (at least not before the last periods) to cheat on trusting investors because of the loss of reputation involved, and can not reap the exorbitant profits of this strategy.

## 3.3 Regression Analysis

We estimate an OLS-regression model of the trustees' back-transfer decision. Since the absolute amount is dependent on the investment received, we use the relative back-transfer as dependent variable, defined as the fraction of points returned over points received (i. e., the quadrupled investment). This means that the "nothing" option translates into 0% of the points that the trustee received are returned; when "compensating", the trustee transfers 25% of the amount she received from the trustor; finally, when choosing "equalize", the back-transfer is 62.5% of the invested amount. Column (1) in table 4 reports the result of a regression that takes the average relative back-transfer per matching group as regressand and as regressors the average investment per matching group and dummies for eyespots and explicit

Tab. 4: OLS-Regression, standard errors in parentheses

| dep. var.:<br>(mean) back-transf. | (1) Obs. = MG | (2) Obs. = subj. | (3) Obs. = dec. |
|---|---|---|---|
| Implicit | 0.003 | –0.003 | –0.005 |
| | (0.027) | (0.023) | (0.024) |
| Explicit | 0.176*** | 0.162*** | 0.255*** |
| | (0.035) | (0.027) | (0.023) |
| Fairness | –0.044° | –0.037** | –0.038** |
| | (0.022) | (0.010) | (0.011) |
| Machiavelli | 0.047 | –0.029° | –0.030° |
| | (0.037) | (0.016) | (0.017) |
| Mean Investment | 0.033*** | 0.034*** | |
| | (0.008) | (0.006) | |
| Investment = 4 | | | 0.064* |
| | | | (0.024) |
| Investment = 7 | | | 0.155*** |
| | | | (0.030) |
| Investment = 10 | | | 0.161*** |
| | | | (0.027) |
| ExplicitXlast3 | | | –0.203*** |
| | | | (0.031) |
| Period | | | –0.005** |
| | | | (0.001) |
| Constant | 0.173° | 0.139** | 0.241*** |
| | (0.092) | (0.047) | (0.039) |
| Adj. $R^2$ | 0.862 | 0.447 | 0.280 |
| Observations | 24 | 144 | 1440 |

° p<0.1, * p<0.05, ** p<0.01, *** p<0.001

treatment as well as the score of the MACH-IV and the average response to the fairness question (high answer indicates low fairness norm). This is the most conservative regression: each matching group represents one independent observation, giving 18 observations in total. We find a highly significant positive effect of the investment level and explicit reputation. On the other hand, we do not find an effect of eyespots on the average relative back-transfer on any imaginable level of significance ($p = 0.898$). The fairness question is weakly significant and the coefficient has the expected sign[20]. This model is able to account for about 86 percent of the observed variation in behavior aggregated for all periods and over matching groups.

Relaxing this most conservative stance in model (2), we examine mean relative back-transfer on the individual level. This yields 144 observations, as we have 36 trustees in the Explicit treatment and 54 in each of the other two. We include the same regressands and get virtually the same results, both concerning size and significance of the coefficients. Fairness is more significant and machiavellianism gets weakly significant (at the 10% level), with a rather small coefficient. More machiavellian participants give back less on average. $R^2$ goes down since we de-aggregate the data.

Moving down further in model (3), we take individual decisions as units of observation, clustering variance on the level of matching groups. The dependent variable is now the individual relative back-transfer, i. e., 0%, 25% or 62.5% of the received amount[21]. As regressands, we are now able to include dummies for the different investment levels instead of mean investment and a dummy that captures the end game effect property of the explicit treatment, as well as a "period" variable because in this regression we can differentiate between periods. The "ExplicitXlast3" dummy is 1 if the observation was made in the explicit treatment during the final three periods and zero otherwise. The baseline in this model is a trustee in the control treatment in whom the trustor invests 1 point. We observe that the coefficient for the implicit treatment has the same magnitude as in the previous model and is still not different from zero on any conventional level ($p = 0.798$)[22]. On the other hand, we still see a positive effect of the explicit condition, which is now even bigger than in the other model. In the non-explicit treatments, trustees send back roughly 15% of the received amount; while in the explicit treatment they transfer on average 37%. Why is the coefficient for the explicit treatment larger in the second model? This can be explained by the fact that we

[20] Concerning the magnitude of the coefficient, a change from "very fair" to "very unfair" would imply a raise in average relative back-transfer of .26; this is about the difference between "nothing" and "compensate".

[21] This means that we have three possible values. Consistency of OLS regressions even with binary cardinal regressands has been shown, see Moffitt (1999)

[22] also, note that the coefficient has the opposite sign than what the subtle cues hypothesis would predict

now differentiate between a basic treatment effect and a treatment effect on end game behavior. The end game coefficient shows that the positive effect during the experiment decreases dramatically in the last periods (cf. figures 2 and 3 which show this end game behavior nicely). As for the investment levels, the more detailed look shows that it is mostly an investment of more than half of the endowment (i. e., 7 or 10 points) that leads to highly increased back-transfers. This means that trustees tend to reward trust by increased returns to the investors. The third model can account for about a fourth of the variation in the individual decisions.

We examined the effect of socio-economic variables that we collected via the questionnaire; for example, we were expecting a gender effect, but neither a gender dummy nor a gender/implicit interaction dummy turned out to be significant. Alltogether, socio-economic variables are jointly insignificant.

## 4 Conclusion

We tried to disentangle two alternative explanations for the empirical observation of strong reciprocity. In order to achieve this goal, we conducted an experiment of repeated one-shot trust games with and without subtle observational cues and compared second mover behavior with a reputational treatment. Our findings suggest that subtle cues do not increase prosocial behavior, at least not in a mutual gift-exchange setting. Moreover, we show that explicit reputational incentives create significantly different trustee behavior than treatments that do not involve reputation forming, including the subtle cues treatment. The former entails rational cooperation of selfish subjects. Since all the rational selfish subjects choose high back-transfers, the overall back-transfer level is greater than in the latter and has a distinctive end game effect. In the non-reputational treatments, selfish subjects do not cooperate while there are other trustees that cooperate and do *not* display end game defection. This indicates that the implicit cues hypothesis cannot account for strongly reciprocal behavior.

These results have important implications for employers. They face the problem that many actions of employees are unobservable or unenforcable. The hidden action problem leaves room for employee shirking. It would therefore be desirable to have a means for mitigating selfish behavior that is detrimental on the firm. Our findings suggest that employees are unlikely to be influenced by subtle cues of being watched. Thus, applying such cues will not be effective in this respect, and are certainly no substitute for real reputational incentives when those are not inplementable.

An open question is how the different regimes with and without reputational incentives are perceived by agents. It may be that cooperative types who were generous throughout the experiment in the baseline condition, turn into rational cooperators as they do not see cooperation as voluntary but enforced by the reputational device. This may crowd out voluntary cooperation as indicated in experiments by Fehr and Gächter (2002b) and Irlenbusch and Sliwka (2005). In order to investigate this point, it would be desirable to conduct additional sessions featuring baseline and explicit in a within-subject design.

## References

Andreoni, J. and J. Miller (1993). Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence. *The Economic Journal 103*(418), 570–585.

Barkow, J. H., L. Cosmides, and J. Tooby (1992). *The Adapted Mind: Evolutionalry Psychology and the Generation of Culture.* Oxford University Press.

Bateson, M., D. Nettle, and G. Roberts (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters 12*, 412–414.

Bolton, G. and A. Ockenfels (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review 90*(1), 166–193.

Brown, M. and C. Zehnder (2005, May). Credit Registries, Relationship Banking and Loan Repayment. Working Paper 240, IEW, Zürich.

Charness, G. and M. Rabin (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics 117*(3), 817–869.

Christie, R. and F. C. Geis (1970). *Studies in Machiavellianism.* New York: Academic Press.

Dufwenberg, M. and G. Kirchsteiger (2004). A theory of sequential reciprocity. *Games and Economic Behavior 47*(2), 268–298.

Falk, A. and U. Fischbacher (2006). A theory of reciprocity. *Games and Economic Behavior 54*(2), 293–315.

Falk, A. and M. Kosfeld (2006). The Hidden Costs of Control. *American Economic Review 96*, 1611–1630.

Fehr, E. and U. Fischbacher (2003). The nature of human altruism. *Nature 425*(6960), 785–791.

Fehr, E., U. Fischbacher, and S. Gächter (2003). STRONG RECIPROCITY, HUMAN COOPERATION, AND THE ENFORCEMENT OF SOCIAL NORMS. *Human Nature 13*(1), 1–25.

Fehr, E. and S. Gächter (2002a). Altruistic punishment in humans. *Nature 415*(6868), 137–140.

Fehr, E. and S. Gächter (2002b, April). Do incentive contracts undermine voluntary cooperation? Working Paper 34, IEW, Zürich.

Fehr, E. and J. Henrich (2003). Is strong reciprocity a maladaptation? on the evolutionary foundations of human altruism. In P. Hammerstein (Ed.), *Genetic and Cultural Evolution of Cooperation*. Cambridge, MA: MIT Press.

Fehr, E. and K. Schmidt (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics 114*(3), 817–868.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Gächter, S. and A. Falk (2002). Reputation and Reciprocity: Consequences for the Labour Relation. *Scandinavian Journal of Economics 104*(1), 1–27.

Gintis, H. (2000). Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology 206*(2), 169–179.

Haley, K. and D. Fessler (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior 26*(3), 245–56.

Hardy, C. and M. Van Vugt (2006). Nice Guys Finish First: The Competitive Altruism Hypothesis. *Personality and Social Psychology Bulletin 32*(10), 1402.

Heintz, C. (2005). The ecological rationality of strategic cognition. *Behavioral and Brain Sciences 28*(06), 825–826.

Irlenbusch, B. and D. Sliwka (2005, September). Incentives, decision frames, and motivation crowding out – an experimental investigation. Discussion Paper 1758, IZA, Bonn.

Johnson, D., P. Stopka, and S. Knights (2003). Sociology: The puzzle of human cooperation. *Nature 421*(6926), 911–2.

Kreps, D., P. Milgrom, J. Roberts, and R. Wilson (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *J. Econ. Theory 27*, 245–252.

Livingston, J. (2005). How Valuable is a Good Reputation? A Sample Selection Model of Internet Auctions. *The Review of Economics and Statistics 87*(3), 453–365.

Moffitt, R. A. (1999). New developments in econometric methods for labor market analysis. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3, pp. 1367–1397. North-Holland.

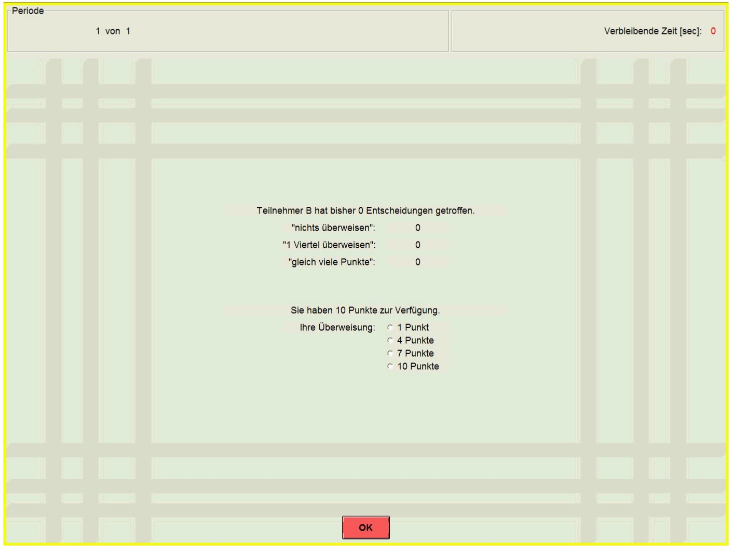
Resnick, P., R. Zeckhauser, J. Swanson, and K. Lockwood (2006). The value of reputation on eBay: A controlled experiment. *Experimental Economics 9*(2), 79–101.

Roth, A. (1995). Bargaining experiments. In J. Kagel and A. Roth (Eds.), *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Trivers, R. (1971). The Evolution of Reciprocal Altruism. *Quarterly Review of Biology 46*(1), 35.

# A Screenshots

Fig. 7: Trustors’ decision screen, featuring the control background



Periode

1 von 1

Verbleibende Zeit [sec]: 0

Teilnehmer B hat bisher 0 Entscheidungen getroffen.

"nichts überweisen": 0

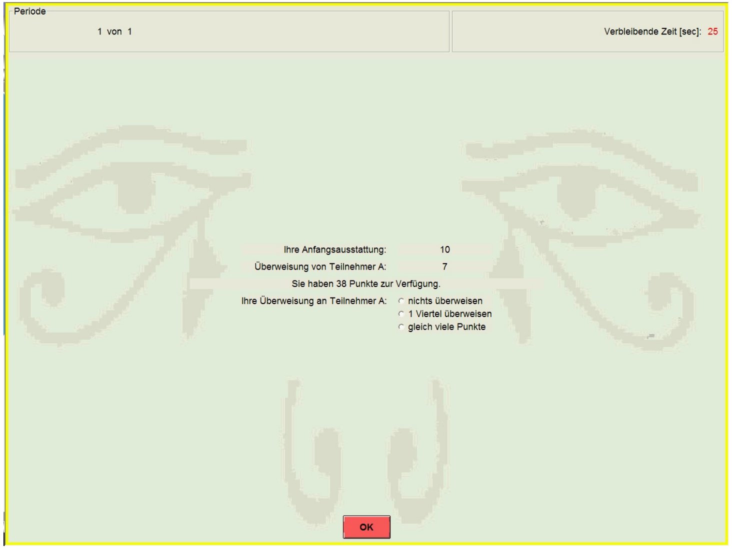
"1 Viertel überweisen": 0

"gleich viele Punkte": 0

Sie haben 10 Punkte zur Verfügung.

Ihre Überweisung:
○ 1 Punkt
○ 4 Punkte
○ 7 Punkte
○ 10 Punkte

Fig. 8: Trustees’ decision screen, featuring the eyespots background



Periode

1 von 1

Verbleibende Zeit [sec]: 25

Ihre Anfangsausstattung: 10

Überweisung von Teilnehmer A: 7

Sie haben 38 Punkte zur Verfügung.

Ihre Überweisung an Teilnehmer A:
○ nichts überweisen
○ 1 Viertel überweisen
○ gleich viele Punkte

# B Tables

Tab. 5: Kolmogorov-Smirnov Tests on Emotion Induction
Difference in Response Distributions between Baseline and Implicit

| Emotion | KS p-value | MWU p-value | Δ in means[a] |
|---|---:|---:|---:|
| "afraid" | 1.000 | 0.9714 | 0.0556 |
| "amused" | 0.965 | 0.9183 | 0.1667 |
| "angry" | 1.000 | 0.9951 | –0.1944 |
| "blue" | 1.000 | 0.9612 | –0.0278 |
| "bored" | 0.615 | 0.6450 | 0.4167 |
| "cheerful" | 0.825 | 0.3895 | –0.3056 |
| "depressed" | 0.615 | 0.9898 | –0.1389 |
| "disgusted" | 1.000 | 0.2628 | –0.4444 |
| "fearful" | 0.965 | 0.5838 | 0.0278 |
| "furious" | 1.000 | 0.1407 | –0.4444 |
| "happy" | 0.413 | 0.4704 | 0.3611 |
| "indifferent" | 1.000 | 0.4908 | 0.4167 |
| "mad" | 0.825 | 0.3581 | –0.1944 |
| "nauseated" | 0.825 | 0.1664 | –0.1944 |
| "nervous" | 1.000 | 0.5547 | 0.1944 |
| "neutral" | 0.615 | 0.1005 | –0.5278 |
| "repulsed" | 0.965 | 0.3297 | –0.1111 |
| "sad" | 1.000 | 0.9143 | –0.1667 |
| "unemotional" | 0.965 | 0.7603 | 0.1944 |

[a] Mean response in Baseline condition – mean response in Eyespots condition, in points (total range of scale: 9 points)