Fehrler, Sebastian; Kosfeld, Michael

Conference Paper

Can you trust the good guys? A test of two theories of reciprocity

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Social Preferences, No. E13-V4

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Fehrler, Sebastian; Kosfeld, Michael (2010) : Can you trust the good guys? A test of two theories of reciprocity, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Social Preferences, No. E13-V4, Verein für Socialpolitik, Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/37173

Can you trust the good guys? A test of two theories of reciprocity

by Sebastian Fehrler* and Michael Kosfeld**

February 2010, Zurich

...we receive a much greater satisfaction from the approbation of those, whom we ourselves esteem and approve of, than those, whom we hate and despise.

David Hume (1896, volume 2, part 1, section 11; first published in 1739)

Abstract

The strive for social esteem is an important motive for pro-social behavior. Many people want to be seen as nice. Recent theories have suggested that the valuation of such esteem depends on the audience. In this study we look at trust and trustworthiness towards people who do or do not identify themselves with typical altruistic goals. In a trust game we observe strong discrimination against those who do but no positive discrimination of those who do not. Moreover, we find that only those second movers who identify themselves with the goals discriminate between the audiences. The last two findings cannot be explained by existing theories.

*University of Zurich, IZA, and Center for Comparative and International Studies (CIS), Email: sebastian.fehrler@pw.uzh.ch.

**University of Frankfurt and IZA.

1 Introduction

In many situations we have to trust people we do not know much about. Can we infer something about a person's trustworthiness from her identification with some altruistic goals, possibly revealed by an Amnesty International T-shirt or a "Save the Whales" badge on the car? Recent theories suggest that this could depend to a large degree on what that other person herself thinks about us.

The strive for social esteem has long been emphasized as an important motive for prosocial behaviour by classical thinkers as David Hume or Adam Smith (see discussion in Ellingsen and Johannesson 2008). Newer theories of prosocial behaviour have taken up their ideas. In the theories by Bénabou and Tirole (2006) and by Ellingsen and Johannesson (2008) the motive of social reputation is central for acting prosocially. The latter theory adds the idea that the valuation of such esteem depends on the audience, an idea which makes it to some extent similar to Levine's theory of altruism and spitefulness (Levine 1998). In this theory, an agent's utility from another agent's payoff depends on the (expected) degree of altruism of the other agent.

Both theories explain reciprocal behavior since kind behavior of another player raises my beliefs about his degree of altruism and makes me, therefore, value his payoff or esteem higher. Other theories of reciprocity, following Rabin (1993), model reciprocity by letting an agent's utility depend on her perception of other players' actions as fair or unfair. Dufwenberg and Kirchsteiger's (2004), and Falk and Fischbacher's (2006) models extend Rabin's theory. In these theories kindness depends on actions, available actions, and beliefs of the players. We agree with Ellingsen and Johannesson's (2008) that it is easier to model kindness as a property of players' preferences, such as their degree of altruism, rather than as properties of actions. Moreover, both Ellingsen and Johannesson's and Levine's theories are better able to explain indirect reciprocity. Indirect reciprocity means that A gives to B because A observed B giving to C. Such behaviour was, for example, observed in laboratory experiments by Dufwenberg et al. (2001) and Albert et al. (2007).

However, both Elligsen and Johannesson's and Levine's theories also predict that all people discriminate between different audiences and predict both positive and negative discrimination of good and bad audiences, respectively. The main point of this study is to test these predictions.

There is evidence from an experiment by Albert et al. (2007) which makes us doubt that the prediction that all subjects discriminate between different audiences holds. Albert et al. study cooperative behavior in a prisoners' dilemma and a dichotomous trust game of subjects

who could donate money to an NGO of their choice in an earlier experiment. They observe that subjects cooperate more often with subjects who donated more money. This is in line with the theories. However, they also observe that subjects who donated little themselves did not show different cooperation levels with different types of other players.

The Albert et al. study itself, however, is not a real test the prediction (this was not the authors' goal) since behavior of both players in the prisoners' dilemma game and of trustors in the trust game does not only depend on their preferences about outcomes for different audiences but also on their beliefs about the other player's behavior. Both can be influenced by an information about the other player's kindness and the two channels are not not studied separately. In our design, we exclude beliefs about the other player's behavior as a motive.

The second prediction to see both negative and positive discrimination comes from the models' feature that subjects simply update their beliefs upon receiving an information about the other player. Therefore, both positive and negative information influences the beliefs and should, therefore, lead to positive and negative discrimination, respectively. Both positive and negative reciprocity have been observed in many experiments (see, e.g., Cox and Deck 2005). We test if positive or negative discrimination of different audiences is more prevalent or if both occur as the theories predict.

To do so, we study trustworthiness towards different audiences. As audiences we consider subjects who do or do not identify themselves with typical altruistic goals like the goals of Amnesty International (AI) and the World Wildlife Fund (WWF). The idea behind this design is that knowing whether or not the other player identifies herself with typical altruistic goals (human rights, environmental protection) should change my belief about the other player's altruism. In a simple trust game subjects can make their transfer decisions conditional on this information about the other player which is elicited with a short survey before the experiment. Trustors' beliefs and transfers indicate whether their beliefs are, indeed, changed by the information. In a control treatment designed to control for mere in-group effects, subjects can condition their transfers on the art preferences of the other subject.

We are interested in the following questions. Are the good guys (those who identify themselves with the goals of AI or the WWF) really more trustworthy (more kind) than average? Do (all) subjects discriminate between kind (good) and less kind (bad) subjects, and if there is discrimination, is it driven by negative or by positive discrimination? How does trust, i.e., trustor behavior, depend on the type of the trustor and the type of the trustee?

While in Levine's (1998) theory esteem does not play a role it is the key element of

Ellingsen and Johannesson's (2008) model. A number of field and laboratory studies have revealed that people show more prosocial activity in public than in anonymous settings. Gächter and Fehr (1999) show that relaxing anonymity can increase contributions in a public good game. Ariely et al. (2009) demonstrate the importance of social approval for charity in a laboratory and in a field experiment. Futher experimental evidence pointing in the same direction is presented by Andreoni and Petrie (2004), Rege and Telle (2004), and Soetevent (2005). In all these studies, it appears that agents value social esteem and expect their prosocial behavior to be esteemed of by the audience.

As an example for experiments in which audience dependent valuation of esteem plays a role, Ellingsen and Johannesson (2008) discuss an experiment by Falk and Kosfeld (2006) which demonstrates the detrimental effect of controlling an agent's actions by restricting her choice set. This signals lacking trust and makes the controlling principal a worse audience for the agent than a trusting principal. As a consequence, the outcomes for a controlling principal are worse even though the control mechanism is available at no cost. Strong negative responses to unkind behavior have been observed in other experiments as well. Fehr and Gächter (2000), for example, show a strong willingness to punish free riders in public good games. Fehr and List (2004) and Fehr and Rockenbach (2003) show the negative effect of choosing contracts with possible sanctions on trustworthiness.

Interestingly, all the experimental studies Ellingsen and Johannesson (2008) present as evidence for their theory are lab experiments in which the subjects interact anonymously. One might ask oneself, whether esteem really has a role in such a context. The authors claim it does and present an experiment by Dana et al. (2006) as evidence. In one treatment of the experiment, subjects playing the dictator role in a dictator game in which they have to split 10 USD can, instead of playing the game, decide to exit and keep 9 USD. In case of exit, the other player never learned that the first player had the choice to play a dictator game, whereas in case of no exit the second player was informed about the game and the first player's decision how to split the pie. In another treatment, the second player was also not told where the money comes from in the second case (where the first player decided not to exit and to make a decision in the dictator game instead). While the authors observe almost no exit in the second treatment they do observe significant exit in the first, indicating that subjects not only care about the other player's payoff. (In this case nobody would exit as allocations (10,0) and (9,1) are feasible in the dictator game and choosing to play the game, therefore, dominates the (9,0) exit option.) Even in the anonymous lab setting subjects apparently care about the thoughts (or esteem) of the other player.

Even if one is still not convinced that esteem plays a role even in anonymous settings, Levine's (1998) theory in which esteem plays no role makes the same predictions.

Another important issue when talking about behavior of members of different groups is the "minimal group paradigm". The term denotes discrimination against people outside one's own group even if the group is arbitrarily formed, a well known phenomenon in social psychology.[1] Recent studies addressing this issue are Chen and Li (2009) and Charness et al. (2007). We control for possible minimal group effects with our control treatment.

We report the following main results. The good guys, i.e., the second movers who identify themselves with the goals of either NGO, are indeed on average significantly more trustworthy. However, they strongly discriminate between the audiences and transfer back substantially more to first movers who identify themselves with either NGO as well, i.e., to the good audience. The comparison to our control treatment shows, that the difference in the backtransfer levels entirely stems from negative discrimination against the bad audience and not from positive discrimination of the good audience. Moreover, we find that second movers who do not identify themselves with the NGO goals do not discriminate at all. The last two findings contradict Ellingsen and Johannesson's and Levine's theories.

The paper proceeds with the experimental design in Section 2, a detailed presentation of the results in Section 3, and the conclusion in Section 4.

2 Experimental Design

We now turn to the experimental design and make a number of predictions of the outcomes of our experiment.

Trust Game

The subjects play a standard trust game. Half of the subjects are trustors the other half trustees. All recipients receive an initial endowment of 12 points. Trustors can transfer 0, 4, 8 or 12 points to the trustee. The transfers are tripled. The trustees can then send back any integer amount of points from the points they have back to the trustor. After the backtransfers, the experiment ends and the subjects are paid out. The experiment consists of only one round.

[1] A classic study is Tajfel et al. (1971). See Crisp and Turner (2007) for a textbook presentation, or Brewer (1979) and Mullen et al. (1992) for reviews of empirical studies on the minimal group paradigm.

In the beginning, before distributing instructions for the trust game, the subjects are asked to fill out a short questionnaire on their computer screens. The questionnaire includes questions like "Do you do sports?", "Do you play an instrument?", and the question "Do you strongly identify yourself with the goals of one of the NGOs, Amnesty International or the WWF?". The last question is the one we are interested in in our main treatment. It has the following answer options: "WWF", "Amnesty International" and "None of the two". One answer option has to be checked, and multiple answers are ruled out.

In the control group setting we use a different question from the same questionnaire to form groups: "Do you very much like one of the painters: Paul Klee or Wassily Kandinski?" with answer options, "Klee", "Kandinsky" and "None of the two". This setting is designed to control for mere in-group effects. We relate to the classic social psychology study in this field by Tajfel et al. (1971) in which preferences about Klee and Kandinski are used as well to form "minimal" groups. As these art preferences do not carry any information about prosociality, no differences between the transfer and backtransfer levels to different subjects should be observed above a possible minimal group effect. The questionnaire is designed to give the subjects the impression that they take part in a small socioeconomic survey. This makes it unlikely that they expect that their answers play a role in the experiment.

In the trust game trustors and trustees can make their transfer decision conditional on the type of the recipient, i.e., on the answer of their partner to the NGO question in the main treatment and on the answer to the art question in the control treatment.

Procedural Details

The trust game is played with the strategy method. Trustors make three transfer decisions, one for each potential type of trustee. Trustees make twelve decisions, one for each possible trustor type and received transfer.

One point in the trust game is worth 0.8 CHF. Overall, 190 subjects participated in the experiment in the laboratory of the Institut für empirische Wirtschaftsforschung (IEW), at the University of Zurich.[2]

Theoretical Predictions

From the models by Ellingsen and Johannesson (2008) and Levine (1998) we now derive predictions for our experiment. In Ellingsen and Johannesson's model agents derive utility

[2]The treatments were programmed with zTree (see Fischbacher 2007).

from their own payoff, from the other player's payoff, and from the pride they take in the other player's esteem of their prosociality. They specifically consider altruism but other forms of prosociality can be modeled similarly (as demonstrated in an earlier version of their paper, Ellingsen and Johannesson 2006).

In their model, which we slightly simplify here (by not considering the case that agents can have biased beliefs in the sense that they expect other people to be like them, which does not influence our predictions), agents maximize the following utility function

$$u_i = m_i + \theta_i m_j + \widehat{\theta}_{ji}$$

with m_i denoting player i's material payoff, θ_i the degree of altruism, with $1 > \theta_i > 0$, and $\widehat{\theta}_{ji}$ player i's pride, defined as

$$\widehat{\theta}_{ji} = E_{\theta_j}[\sigma(\theta_j)\theta_{ji}]$$

with $\sigma(\theta_j)$ being the salience of the opponent's esteem and θ_{ji} the esteem of player j of player i, defined as

$$\theta_{ji} = E[\theta_i|h]$$

with h denoting the history of the game.

The valuation of esteem, $\sigma(\theta_j)$, depends on the expected degree of altruism of the other player. It is assumed that σ is increasing in θ_j, i.e., esteem from more altruistic agents is valued higher.

In Levine's (1998) theory the utility of an agent in a two player game is given by the following function

$$u_i = m_i + (\theta_i + \lambda\theta_j)m_j$$

where m and θ have the same meanings as above and $\lambda \geq 0$ is a parameter common to all players.

Both theories predict that kind people are treated better (if $\widehat{\theta}$ or λ are greater than zero). They also predict that all subjects discriminate between good and bad audiences (neither the valuation of the esteem nor the valuation of the other player's altruism depends on the player's own type). Moreover, both theories predict that kind people are treated better than average kind people and less kind people treated worse.

In our setting kind behavior takes the form of high backtransfers (relative to the trustor's transfer) or trustworthiness.

As we compare backtransfers in the trust game for each transfer level, the history h of the game does not play a role. The other factor plausibly influencing the expectation of the other player's type is the information on her answer to the NGO question. We can think of this information as also being represented by the term h.

We predict that subjects show higher trustworthiness towards a good audience, i.e., towards subjects who identify themselves with either NGO. We predict to see this pattern for all trustees. Moreover, as we expect that identification with either NGO increases the expected θ of the other player as well as the contrary information decreases it, we predict to see higher (lower) trustworthiness towards the good (bad) audience as compared to trustworthiness in the control treatment.

We also expect trustors to trust kind trustees more than unkind trustees and, therefore, transfer higher amounts to them. This is crucial as it is a test of the assumption that the good (kind) guys are, indeed, perceived as a better audience.

3 Results

We now turn to the presentation of the results. Of the 190 participants there are 32% subjects who identify themselves strongly with the goals of the WWF, 26% with the goals of Amnesty, and 42% with neither NGO's goals. There are 29% subjects who state to like Klee, 22% to like Kandinski, and 49% who indicate to like neither painter.

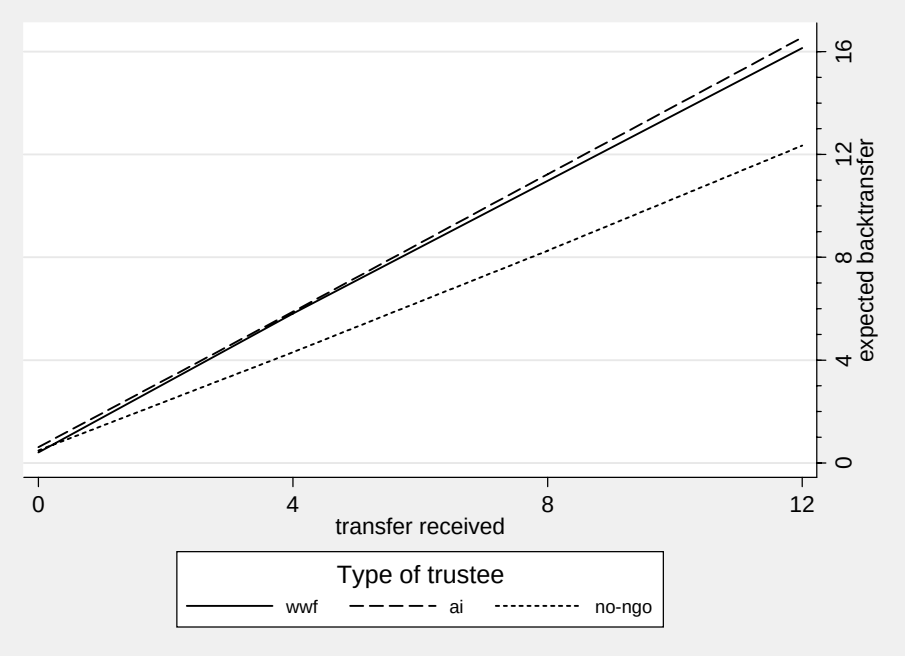
Trustor Behaviour

We asked the trustors about their beliefs regarding backtransfers for all possible transfer levels and types of trustees. In Figure 1 we see that trustors expect lower backtransfers from subjects who do not strongly identify themselves with the goals of either NGO, henceforth called No-NGO types. Looking at different types of trustors separately, shows that this is true for all types of trustors (see Figure 5 in the Appendix).

Moreover, we see that the beliefs about backtransfers from AI or WWF types are almost the same. Table 1 shows the transfer levels to the different types of trustees by the different types of trustors. The differences between the transfer levels reflect the beliefs about the backtransfers. Even the No-NGO types transfer less to other No-NGO types than to AI or WWF types. For the No-NGO types, the differences of the transfer levels to the three trustee types are pairwise statistically different at the 5% level (Wilcoxon rank sum test).[3] For the

[3]All test results we report are for undirected hypotheses.

Figure 1: Beliefs about the trustworthiness of different trustees



other two groups the transfer level to No-NGO types is statistically different at the 5% level from the other two groups which themselves are not significantly different from each other. Transfers to No-NGO types are lower than to any other group. The NGO types receive, on average, 47% higher transfers than No-NGO types. This shows that NGO types are believed to be more trustworthy (i.e., to have a higher θ) than No-NGO types. Our assumption that information about a player's identification with altruistic goals changes beliefs about her kindness is thus confirmed. NGO type trustors will, therefore, also be a better audience for trustees.

Table 2 shows the transfer levels in the control group. Here, each type of trustor favors trustees with the same art preferences but there is no group nobody trusts less than all other groups.

Trustee Behavior

In the analysis of trustee behavior we start by looking at the trustworthiness of the different NGO types in the control group setting where they cannot condition their backtransfer on the NGO type of the trustor. This allows us to see whether the NGO types are more trustworthy in general. In the control treatment the transfers have to be conditioned on

Table 1: Transfer levels from different NGO types to different NGO types[a]

Transfer	to WWF	to AI	to No-NGO	N
from WWF	8.3 (0.8)	7.9 (0.8)	4.6 (0.9)	28
from AI	7.1 (1.2)	8.3 (1.2)	4.6 (1.8)	14
from No-NGO	6.8 (0.7)	7.6 (0.7)	5.9 (0.8)	36

[a] We use NGO type and then just the NGO name as abbreviations for subjects strongly identifying themselves with the goals of that NGO. Standard errors in parentheses.

Table 2: Transfer levels from different artist types to different artist types[a]

Transfer	to Klee	to Kandinski	to No-Artist	N
from Klee	8.6 (1.0)	6.6 (1.1)	6.4 (1.1)	22
from Kandinski	7.6 (0.9)	8.8 (0.8)	6.7 (1.1)	19
from No-Artist	6.3 (1.0)	6.2 (1.0)	8.3 (0.9)	26

[a] We use artist type and then just the artist's name as abbreviations for subjects liking the work of the artist a lot. Standard errors in parentheses.

the art preferences of the trustor. As we used the same questionnaire for both control and treatment group we can group the control group results by the answers to the NGO question. This allows us to study trustworthiness of the different NGO types in a neutral setting (the artist type of the other recipient should not play a role since it is an irrelevant information regarding expected trustworthiness as the analysis of trustor behavior showed). Figure 2 presents the backtransfers for the different potential transfers averaged over the three potential recipient types ("Klee", "Kandinski", and "No-Artist").

We see that people who identify themselves with one of the NGOs are more trustworthy than people who do not, just as trustors expect. Regressing backtransfer on transfer gives significantly different slopes for the AI group than for the No-NGO group (at the 5% level). In this regression there are four observations from every trustor (one for each possible transfer level). This is taken into account in the estimation of the standard errors by treating these four observations as one cluster each. The difference between the slopes is tested using an adjusted Wald test. Pooling the AI and the WWF group in the regression gives a significantly

Figure 2: Actual trustworthiness of different NGO types[a]

[a] Average backtransfers from the control treatment in which transfers could only be made conditional on the art preferences of the receiver. N=67.

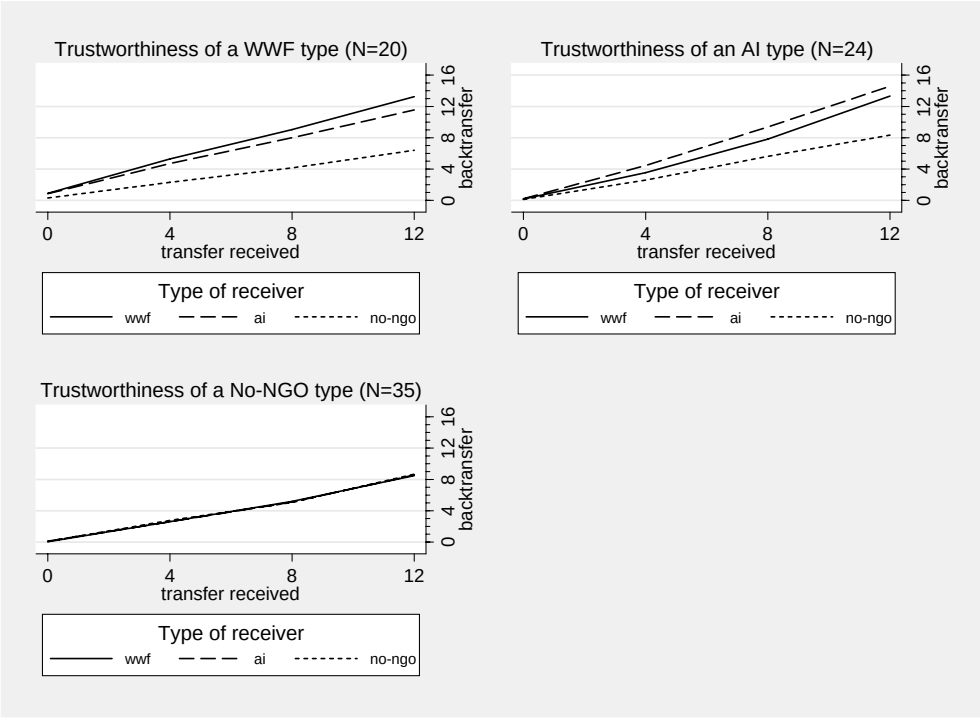
different slope of this combined NGO group to the No-NGO group slope (at the 5% level). The backtransfer after a transfer of 12 is also significantly different between the WWF and the No-NGO group (at the 10% level, Wilcoxon rank sum test). The slope of the WWF group alone is not significantly different to the slope of the No-NGO group. It is also not significantly different to the slope of the AI group.

This tells us that subjects who identify themselves with an NGO are, indeed, more kind in the sense of being more trustworthy. The backtransfers they make are about 1.5 times higher than the backtransfers of the No-NGO types.

Do trustees discriminate when they face different audiences? An answer to this question is given by the three graphs in Figure 3.

We see that NGO types, the good guys, do, indeed, strongly discriminate against No-NGO types. The slopes of the regression lines when regressing backtransfer on transfer by trustor type are significantly different from each other in case of an Amnesty and an WWF trustee (at the 5% level, adjusted Wald tests, standard errors clustered as above). It is also the case that AI types favour other AI types over WWF types, and WWF types favour other WWF types over AI types. This can be explained by the mere in-group effect which is present

Figure 3: Trustworthiness of different NGO types towards different trustors

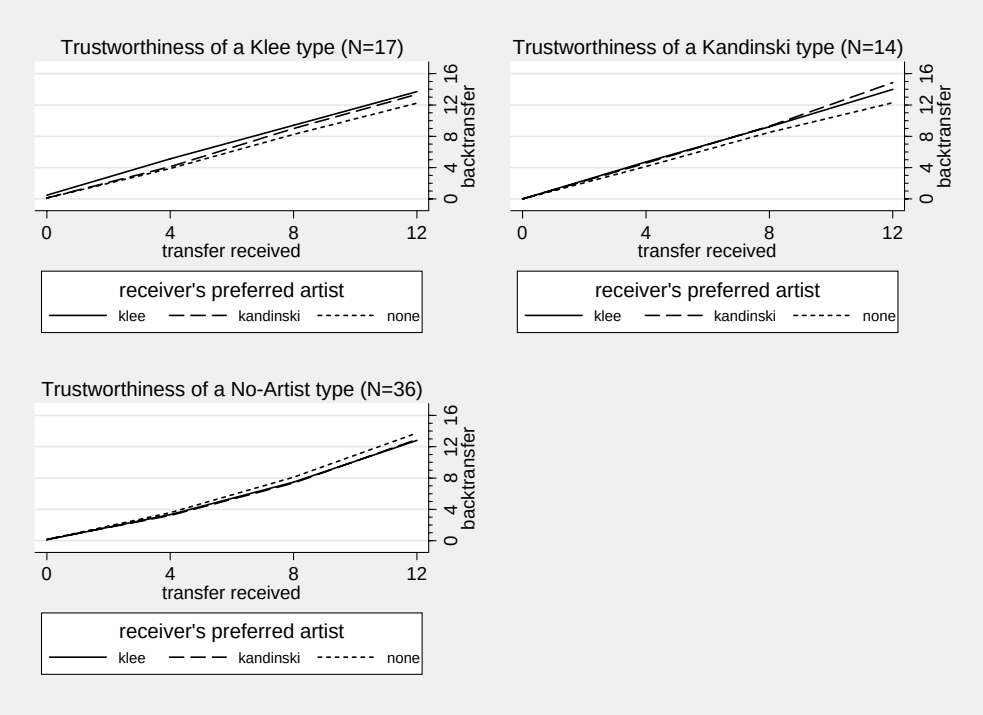


even when group formation is arbitrary as in our control treatment. Figure 4 shows that there are small differences between the artist types at high transfer levels. These differences are tiny, though, compared to the difference between the good and the bad audience group.

Interestingly, we see in the beliefs of the No-NGO trustors in Figure 1 and in their transfers in Table 1 that they did not anticipate to be discriminated against.

Backtransfers to either NGO type trustor from either NGO type trustee are on average 1.8 times higher than backtransfers to No-NGO types. This difference is large and in line with our first theoretical prediction that the different audiences will be treated differently. However, we find that No-NGO types (third panel in Figure 3) do not discriminate between the audiences. This contradicts our prediction to observe the same pattern for all trustees. Subjects who are expected to have a lower θ (and all groups even the No-NGO types themselves expect this from the No-NGO group) and are, indeed, less trustworthy in the neutral setting of the control treatment (Figure 2), apparently do not value esteem from (or in the context of Levine's theory the payoff of) subjects more who have a higher θ. In other words, a selfish type does not care whether the other player is selfish too or not.

Figure 4: Trustworthiness of different Artist types towards different trustors



What drives the discrimination of the different audiences by the NGO types? The comparison of the behavior of the NGO types in the treatment and the control group reveals a strong negative discrimination against the bad audience and no positive discrimination of the good audience (Figures 2 and 3). A regression of backtransfers from either NGO group (we pool the two groups in the regression) to either NGO group on the received transfer gives us two slope parameters. Regressing average backtransfers from either NGO group (we again pool the two groups) to the different artist types on the received transfer gives us another slope coefficient. The latter is statistically not different to the coefficients from the first regression (at the 5% level, adjusted Wald tests, standard errors clustered as above). The slope coefficient in a regression of backtransfers from the pooled NGO group to No-NGO types on the received transfer, however, is significantly lower than the slope for the pooled NGO group in the control group setting (at the 5% level, adjusted Wald test, standard errors clustered as above). These results also hold if NGO groups are not pooled and AI and NGO groups are looked at separately.

This finding contradicts our theoretical prediction to observe both negative and positive

discrimination, because of the changed expectations about the other person's type. What is driving the differences between the observed trustworthiness toward different audiences is clearly negative discrimination of the bad audience.

These findings suggest that the theories of Ellingsen and Johannesson and of Levine should be modified. In the first theory the pride term and in the latter theory the λ term have to be changed to explain the experimental results. The terms apparently depend on the type of the person in whose utility function they appear. The results suggest that people with a low θ should have a lower weight (pride or λ term, respectively), possibly zero, on the type of the other player in their utility function. Moreover, the theories might have to be modified such that the payoff to low θ types is valued much less than the payoff to an average θ type but that the payoff to high θ types is valued only a little higher than the payoff to an average θ type, if at all.

4 Conclusion

We conducted a simple trust experiment to gain some insight into the old idea that the valuation of social esteem depends on the audience. Ellingsen and Johannesson (2008) formalized this idea and we derive predictions from their and Levine's (1998) theories which we test. We form different audience groups on the basis of the subjects' identification with typical altruistic goals. The idea behind this design is that identification with altruistic goals indicates altruism.

In our trust game, subjects can make their decisions dependent on the type of the other player they have to interact with. They are informed about whether the other player identifies herself with the goals of Amnesty International, or the WWF, or none of the two. This information is elicited in a short survey before the experiment.

Our first finding is that the good audience, i.e., the subjects who stated to strongly identify themselves with the goals of an NGO are, indeed, expected to be more trustworthy. This also means that being perceived as good comes with economic benefits in form of higher trust. Prosocial activities like charity could, therefore, have a role as a signaling device for trustworthiness. Fehrler (2009) shows that the observation of higher transfers remains unchanged if groups are build on the basis of a public voluntary donation to Amnesty International.

The main focus of this study is on trustee behavior, and the next question is whether the good guys are, indeed, more kind than the others. We find that subjects identifying

themselves with one of the NGOs are more trustworthy than subjects who do not identify themselves with either NGO and on average transfer back substantially more. However, we also find that the good guys strongly discriminate against the bad audience (trustors who do not identify themselves strongly with either NGO). Backtransfers to the good audience recipients are on average 1.8 times higher.

However, backtransfers to other NGO types are not higher than average backtransfers in the control group setting. This contradicts the prediction from Ellingsen and Johannesson's and Levine's theories to see both positive and negative discrimination because the expectation of the other person's trustworthiness rises or falls with the information on her type.

Another contradiction of the two theories is that the trustees who do not identify themselves with either NGO do not discriminate at all. This contradicts the prediction to see the same pattern for all trustees. Neither Elligsen and Johannesson's nor Levine's theory explain this.

To accommodate for this finding, the pride term in Ellingsen and Johannesson's theory would have to depend on the type of the subject whose pride it describes and not only on the audience's type. In Levine's theory the same applies to the term capturing utility from the other subjects' payoff. Further adjustments have to be made to the utility functions to explain the finding that we only observe negative but no positive discrimination.

Returning to the title question, whether one can trust the good guys, we conclude that if oneself needs to decide whether to trust somebody or not, taking into account what the other person probably thinks about oneself, i.e., taking into account one's own type (irrespectively of one's own potential actions) appears to be important. The subjects in our experiment interestingly do not anticipate any form of discrimination from trustees.

References

Albert, M., W. Güth, E. Kirchler, and B. Maciejovsky (2007): "Are we nice(r) to nice(r) people? An experimental analysis," *Experimental Economics*, 10(1), 53–69.

Andreoni, J., and R. Petrie (2004): "Public goods experiments without confidentiality: a glimpse into fund-raising," *Journal of Public Economics*, 88(7-8), 1605 – 1623.

Ariely, D., A. Bracha, and S. Meier (2009): "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, 99(1), 544–55.

Bénabou, R., and J. Tirole (2006): "Incentives and prosocial behaviour," *American Economic Review*, 96(5), 1652–1678.

Brewer, M. B. (1979): "In-Group Bias in the minimal intergroup situation: A cognitive-motivational analysis," *Psychological Bulletin*, 86(2), 307–324.

Charness, G., L. Rigotti, and A. Rustichini (2007): "Individual Behavior and Group Membership," *American Economic Review*, 97(4), 1340–1352.

Chen, Y., and S. X. Li (2009): "Group Identity and Social Preferences," *American Economic Review*, 99(1), 431–57.

Cox, J. C., and C. A. Deck (2005): "On the Nature of Reciprocal Motives," *Economic Inquiry*, 43(3), 623–635.

Crisp, R. J., and R. N. Turner (2007): *Essential Social Psychology*. London: Sage.

Dana, J., D. M. Cain, and R. M. Dawes (2006): "What you don't know won't hurt me: Costly (but quiet) exit in dictator games," *Organizational Behavior and Human Decision Processes*, 100(2), 193 – 201.

Dufenberg, M., U. Gneezy, W. Güth, and E. V. Demme (2001): "Direct versus Indirect Reciprocity: An Experiment," *Homo Oeconomicus*, 18, 19–30.

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

Ellingsen, T., and M. Johannesson (2006): "Pride and Prejudice: The Human Side of Incentive Theory," CEPR Discussion Paper 5768.

——— (2008): "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98(3), 990–1008.

Falk, A., and M. Kosfeld (2006): "The Hidden Costs of Control," *American Economic Review*, 96(5), 1611–1630.

Fehr, E., and S. Gächter (2000): "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 90(4), 980–994.

Fehr, E., and J. A. List (2004): "The Hidden Costs and Returns of Incentives-Trust and Trustworthiness Among CEOs," *Journal of the European Economic Association*, 2(5), 743–771.

Fehr, E., and B. Rockenbach (2003): "Detrimental effects of sanctions on human altruism," *Nature*, 422(6928), 137–140.

Fehrler, S. (2009): "Prosocial Behavior as a Signal of Trustworthiness," Available at SSRN: http://ssrn.com/abstract=1502565.

Fischbacher, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10(2), 171–178.

Gächter, S., and E. Fehr (1999): "Collective action as a social exchange," *Journal of Economic Behavior & Organization*, 39(4), 341 – 369.

Hume, D. (1896): *A Treatise of Human Nature*. Oxford: Clarendon Press.

Levine, D. K. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, 1, 593–622.

Mullen, B., R. Brown, and C. Smith (1992): "Ingroup bias as a function of salience, relevance, and status: An integration," *European Journal of Social Psychology*, 22(2), 103–122.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83(5), 1281–1302.

Rege, M., and K. Telle (2004): "The impact of social approval and framing on cooperation in public good situations," *Journal of Public Economics*, 88(7-8), 1625 – 1644.

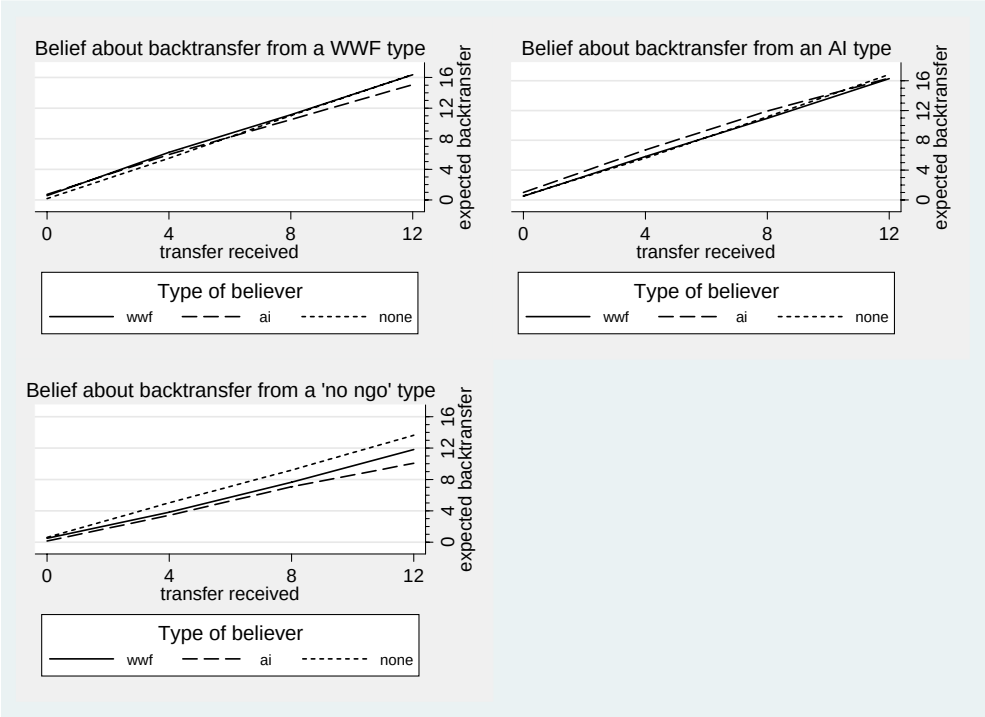
Soetevent, A. R. (2005): "Anonymity in giving in a natural context–a field experiment in 30 churches," *Journal of Public Economics*, 89(11-12), 2301 – 2323.

Tajfel, H., M. Billig, R. Bundy, and C. Flament (1971): "Social Categorization and intergroup behaviour," *European Journal of Social Psychology*, 1(2), 149–178.

Appendix

Figure 5 shows the beliefs of first movers about backtransfers from second movers grouped by the types of first movers.

Figure 5: Beliefs of NGO types about trustworthiness of different second movers (who make the backtransfer)



Regressing the belief about the backtransfer on the first mover's transfer gives us estimates of the slopes of the different lines in Figure 5. The lines for AI and WWF type second movers are significantly steeper than the lines for No-NGO types in all three pictures (at 5% in the first two and at 10% for No-NGO type first movers). In these regressions there are four observations from every first mover, one for each possible transfer level. This is taken into account in the estimation of the standard errors by treating these four observations as one cluster each and using Taylor linearized standard errors. The difference between the slopes is tested using an adjusted Wald test.