Matthey, Astrid; Regner, Tobias

Working Paper

More than outcomes: A cognitive dissonance-based explanation of other-regarding behavior

Jena Economic Research Papers, No. 2011,024

Provided in Cooperation with:
Max Planck Institute of Economics

ECONOMIC RESEARCH PAPERS

2011 – 024

More than outcomes: A cognitive dissonance-based explanation of other-regarding behavior

by

Astrid Matthey
Tobias Regner

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

More than outcomes: A cognitive dissonance-based explanation of other-regarding behavior

Astrid Matthey ♠* Tobias Regner ♠

♠*Max Planck Institute of Economics, Jena, Germany*

May 26, 2011

Abstract

Recent research has cast some doubt on the general validity of outcome-based models of social preferences. We develop a model based on cognitive dissonance that focuses on the importance of self-image. An experiment (a dictator game variant) tests the model.

First, we find that subjects whose choices involve two psychologically inconsistent cognitions indeed report higher levels of experienced conflict and take more time for their decisions (our proxies for cognitive dissonance). Second, we find support for the main model components. An individual's self-image, the sensitivity to cognitive dissonance, and expected behavior of others have a positive effect on other-regarding behavior.

JEL classifications: C72, C91, D03, D80

Keywords: social preferences, other-regarding behavior, self-image, experiments, cognitive dissonance, social norms, normative beliefs, expectations

*Corresponding author (matthey@econ.mpg.de, phone: +49 3641 686644).

We would like to thank audiences at the EEA 2010 congress and at IMEBE 2009 for their feedback. We are grateful to James Konow, Rosemarie Nagel, Ondřej Rydval, Joel Sobel and Toru Suzuki for valuable comments. Michael Enukashvili, Nadine Erdmann and Andreas Lehmann provided excellent research assistance.

1 Introduction

By now other-regarding behavior is an established result in social dilemma and allocation situations. Outcome-based models of social preferences explain these findings with individuals deriving utility from others receiving positive payoffs. However, recent empirical evidence has cast some doubt on this approach. Dictator exit experiments analyzed subjects' behavior when a costly exit option to get out of a dictator game is provided (Dana et al., 2006; Lazear et al., 2006; Broberg et al., 2007). A substantial amount of subjects sorts out of the dictator game, although this means they get a lower payoff. DellaVigna et al. (2009) provide evidence of sorting out in a related field setting. Strategic ignorance experiments analyzed subjects' allocation game choices when they could have avoided being informed about the consequences of their own choice on others (Dana et al., 2007; Larson and Capra, 2009; Grossman, 2010a; Matthey and Regner, 2011). Significantly less subjects behave other-regarding when the consequences can be avoided in comparison to a transparent baseline case with full information.

Hence, alternative approaches to explain other-regarding behavior consider not only own and others' payoffs. The focus of social approval models is on social reputation. In addition to monetary payoffs they allow individuals to be also motivated by the desire to be liked and respected by others (see, for instance, Bénabou and Tirole, 2006; Ellingsen and Johannesson, 2008; and Andreoni and Bernheim, 2009). Social reputation may well matter, but it requires that one's action is signalled to the relevant community. Only part of the empirical evidence mentioned before deals with situations where subjects' decisions are 'public'. When the decision remains 'private' the desire to gain social approval cannot quite explain other-regarding behavior. Hence, it seems sensible to broaden the scope by considering as well a person's self-image and the desire to maintain it.[1] The aim of this paper is to further contribute to the research on the nature of social preferences. Cognitive dissonance serves as the psychological basis to model the effects of self-image on behavior. The paper also provides a detailed experimental test of the

[1] While the focus of Bénabou and Tirole (2006) is on social reputation, they leave room for self-respect, too. They model self-image concerns more explicitly in Bénabou and Tirole (2011). Bodner and Prelec (2003) propose a self-signalling model where actions provide an informative signal to ourselves. In the context of honesty Mazar et al. (2008) develop a theory of self-concept maintenance.

model's hypotheses.

Cognitive dissonance is a psychological theory developed by Leon Festinger (1957). A person experiences cognitive dissonance when she holds two psychologically conflicting cognitions. For example, she may find a certain task boring, but claims that it was interesting as an internal justification for actually doing it (Festinger, 1957). The modern theory of cognitive dissonance (Aronson, 1994; Beauvois and Joule, 1996) argues that dissonance primarily revolves around the self and a piece of behavior that violates that self-concept.[2] Our model is most related to this modern version of dissonance theory. It explains other-regarding behavior as being driven by individuals' desire to maintain their self-image. Any divergence of actual behavior from that self-image would lead to the unpleasant feeling of cognitive dissonance.

In contrast to social psychology where dissonance theory has been frequently applied only few articles in economics use it to explain decision making (see, for instance, Akerlof and Dickens, 1982; Rabin, 1993; and Oxoby, 2003, 2004). In particular, we build on Konow (2000), a detailed cognitive dissonance-based model of other-regarding behavior in dictator games. Our model generalizes Konow's "accountability principle" to allow for subjects to apply different standards of behavior. It waives self-deception as an explicit model component. Instead, it includes an individual's self-image and the sensitivity to cognitive dissonance (the steepness of the dissonance function).

The present experiment is designed to test these model components. Participants first face a standard dictator situation where they allocate an endowment between themselves and an unknown receiver. Then they make three further dictator decisions with the same endowment, knowing that the assigned allocation reaches the receiver only with a probability p (between 80% and 90%). With the remaining probability the dictator keeps the entire endowment and the receiver gets nothing, even if the dictator allocated a positive amount. After all allocations have been decided, subjects learn that they can choose which of the four situations - under the deterministic or stochastic allocation regime - they want to apply. This gives them the opportunity to reduce their expected allocation without derogating their initial (good) intentions.[3]

[2] See Harmon-Jones and Mills (1999) for a review of the current state of dissonance theory.

[3] Rabin (1995) treats people's choices to seek or avoid information regarding the effects of their choices more formally.

The design allows us to i) assess a subject's degree of other-regarding behavior in a standard allocation decision, and ii) get an indication to what extent the subject is willing to yield to the temptation of possibly keeping all the money (by selecting a transfer probability for the allocation of $p < 1$). In addition, it allows us to analyze the factors that influence these decisions.

Our model predicts that pro-social subjects who send equal amounts in the four dictator situations but then choose a situation with $p < 1$ should experience more cognitive dissonance than, for instance, pro-social subjects who send equal amounts and choose the allocation under certainty ($p = 1$). We do find that subjects whose choices indicate that they face a close tradeoff between monetary utility and the disutility from cognitive dissonance report higher levels of experienced conflict and take more time for their decisions (our proxies for the relevance of cognitive dissonance). Moreover, our results indicate strong support for the significance of the model's components: individuals' self-image, how important it is to have and comply with certain principles in life, how unpleasant it is not to comply with these principles, and individuals' beliefs regarding the behavior of others.

The experimental results add another piece of empirical evidence showing that conventional outcome-based models of social preferences fail to explain behavior in certain situations. They suggest that other-regarding behavior to a significant degree is caused by subjects avoiding the cognitive dissonance that would arise if they behaved egoistically although they perceive themselves as pro-social.

The paper is organized as follows. Section 2 develops a simple cognitive-dissonance based model of other-regarding behavior. In section 3 we describe the experimental design and derive testable hypotheses. Results are presented and discussed in section 4. Section 5 concludes.

2 Theory

2.1 Social preferences and cognitive dissonance

Observed other-regarding behavior is usually explained with social preferences. This means that people are assumed to be truly concerned about how much others have (e.g., in the models of Fehr and Schmidt, 1999, and Bolton and Ockenfels, 1999).

However, recent experiments (Konow, 2000; Lazear et al., 2006; Dana et al., 2006; Dana et al. 2007; Broberg et al., 2007; Larson and Capra, 2009; Grossman, 2010a; Matthey and Regner, 2011) suggest that a substantial amount of individuals do not really value equal payoffs. Although they may share with others if forced into an allocation situation, some subjects in these experiments avoided the allocation decision if given the choice, even if this led to reductions in their maximum payoff.

To explain such behavior in a general class of situations, it is helpful to relate it to existing theories in psychology. We follow this approach and model "social preferences" as based on psychological evidence. This is motivated by the observation that other-regarding behavior to a considerable part seems not to be due to genuine social preferences. This is true even if the setup is anonymous and one-shot, i.e. group pressure, reputation, etc. cannot explain the behavior.

The concept we employ to explain such behavior is not new. It was first described as "cognitive dissonance" by Leon Festinger (1957) as the negative drive state that arises if a person holds two cognitions that are psychologically inconsistent. For example, there may be a dissonance between a person's beliefs and her behavior, which is experienced as unpleasant, and produces a motivation to reduce this dissonance. The concept has been used to explain economic behavior, for example by Akerlof and Dickens (1982), Rabin (1994), Konow (2000), and Oxoby (2003, 2004).

In allocation situations, people may have a certain perception about what they should do, what they would like themselves to do, what the person they would like to be would do etc. This is where preferences come in. Rather than having genuine preferences regarding the payoffs of others, people may have a preference over the existence of a certain moral or ethical standard in society, and of their role in forming it, or contribution

to it. For example, they may have a preference for a world where fair behavior is the standard, and acknowledge that if they want such a world, they should behave fairly too. Alternatively, some people may have a preference for a world where it is the agreed standard that everyone behaves in a selfish way, and no one is blamed for that. From their preference for an ethical standard people then derive the picture of the person as that they see themselves, that they would like to be etc.

In this sense, people may have a preference regarding "behavior towards others", and hence also regarding *their own* behavior towards others, but not necessarily regarding these others directly. In the examples above, one may have a preference for people treating each other fairly, but beyond that not care much about the outcomes. In other words, people may have a preference for a *standard of good behavior*, rather than for particular outcomes.

If people behave in a way that deviates from the standard by which they measure themselves, they experience cognitive dissonance (e.g. Aronson, 1994; Beauvois and Joule, 1996). Since this results in unpleasant feelings, they would prefer to avoid such deviations. As we will show below, this reasoning can explain standard other-regarding behavior, but also the behavior we observe in the experiment, which cannot easily be explained by earlier models.

2.2 The model of Konow (2000)

Our model is based on the same general idea as Konow's (2000). Konow was the first to use the concept of cognitive dissonance to explicitly explain behavior in dictator games. In his model, individuals maximize a utility function that is increasing in the share of money they keep for themselves, y, and decreasing in the amount of cognitive dissonance they experience and the extent of self-deception they engage in. Cognitive dissonance arises if the share y an individual keeps for herself deviates from the share she believes is fair to keep, ϕ. Self-deception is present if the individual makes herself believe that a certain share ϕ is fair to keep, but this share deviates from the fair entitlement η. This fair entitlement is determined exogenously, and identically for all individuals and across all situations, by the *accountability principle*: individuals' payoffs should not depend on variables they have no control over.

Given the share y and the belief ϕ, the amount of cognitive dissonance an individual experiences in Konow's model depends on the parameter α, which determines the relevant dissonance function and may vary across individuals and contexts. Similarly, given the fair entitlement η and the individual's belief about the fair share ϕ, the (emotional) costs of self-deception that arise if $\eta < \phi$ depend on the parameter β. The individual then solves the problem

$$\text{Max}_{y,\phi} \quad u(y, \phi, \eta, \alpha, \beta) \equiv v(y) - f(y - \phi, \alpha) - c(\phi - \eta, \beta)$$

that is, she chooses the share to keep and her belief of the fair share. These choices result in the monetary payoff $v(y)$, and potential costs of cognitive dissonance, $f(y - \phi, \alpha)$, and self-deception, $c(\phi - \eta, \beta)$.

Based on Konow's model we generalize his assumption and allow individuals to apply different behavioral standards. This gives the model more flexibility than assuming that everybody agrees on the same behavioral standard, be it derived from one principle as in Konow (2000), or from a set of principles as in his more recent models (Konow 2001, 2010).[4] On the one hand, this seems a more robust assumption given the diverse attitudes people express in real life. On the other hand, it allows the analysis of situations where the accountability principle cannot be applied in a straightforward manner, i.e., where it is not clear what a person's entitlement is.

Moreover, we do not explicitly model self-deception. Rather, self-deception can take two forms in our setup: First, if self-deception is completely successful, i.e., the individual truly believes that ϕ is her "fair share", self-deception does not induce costs since the individual is not aware of it. It therefore does not show up in the utility function. Second, if self-deception is not (completely) successful, i.e., the individual is not fully convinced that ϕ is her fair share, she knows to some degree that choosing ϕ is not fair. Doing so therefore causes cognitive dissonance, i.e., this form of self-deception enters utility through the dissonance function.

[4] In contrast to the current paper, the aim of Konow's (2001, 2010) models is to specify a principle that a majority of people agree with. Arriving at a "one for all" standard is therefore a desirable result in his papers, that does not deny that some people may deviate from it.

2.3 Cognitive dissonance in allocation situations

First we would like to emphasize that this is not meant as a general model of cognitive dissonance, based on all the available psychological evidence. Rather, it is an attempt to explain certain patterns of behavior observed in allocation situations based on the concept of cognitive dissonance.

The setup of our model is the following. Let $\mathbf{X}$ denote the two-dimensional choice set of the individual. Each element $\mathbf{x_i}$ of $\mathbf{X}$ is a vector that contains the individual's allocation to herself, x_i, and the allocation to the other person, x_{-i}.

Individual behavior can be classified into certain standards. Let $\mathbf{\Delta}$ denote a finite set of behavioral standards or rules. An element Δ of $\mathbf{\Delta}$ then denotes a particular behavioral standard, e.g., *fairness*, *cleverness*, *generosity* etc. Each individual has a preference relation $\succeq$ over the set of behavioral standards, where $\succ$ denotes strict preference and $\sim$ denotes indifference. Preference relations are assumed to be complete and transitive.

An individual's standard Δ_i^* reflects her self-image. For example, if an individual prefers to be fair, she has the self-image of being a fair individual. This does not necessarily mean that fairness is *in general* her most preferred standard. For example, overall she may prefer generosity over fairness, but think of fairness as a "sufficient" standard of behavior. Allowing the self-image Δ^* to differ between individuals accounts for the case that one individual prefers fair behavior for herself, while another prefers selfish behavior or generous behavior (individuals have the self-image of being fair, selfish, or generous). This concept includes Konow's (2000) assumption of a universal standard of fair behavior as a special case, but considers the possibility that people differ in their true behavioral preferences (or at least in what they believe are their behavioral preferences without engaging in costly self-deception).

We will assume that in the short run Δ_i^* is constant: people do not change quickly the behavioral standard they prefer for themselves, i.e., the behavior they consider as moral or appropriate. It is not a choice variable in the sense that it can be chosen to optimally suit each new situation.

To connect behavioral standards to choices, we define a set of behavioral standards

$\mathbf{S}(\mathbf{x_i}) \subseteq \mathbf{\Delta}$ that are consistent with a choice $\mathbf{x_i}$.[5] For example, keeping everything for oneself in a standard dictator game is consistent with the standards of *rationality*, *spite*, *efficiency* etc., while allocating 50% to the other player is consistent with *fairness* and *efficiency*. If a choice $\mathbf{x}$ is consistent with several standards $\Delta(\mathbf{x})$, we will assume that the standard implied by observing $\mathbf{x}$ is the one that among all the standards consistent with the behavior is preferred by the individual. For example, the standards *fairness* and *generosity* may induce the same choice if only few choices are available. For an individual that prefers generosity, if she makes this choice we will judge her behavior as generous. Given this assumption, we can say that any choice $\mathbf{x}$ implies a behavioral standard $\Delta_I(\mathbf{x})$:

Definition 1 *: An individual's choice* $\mathbf{x_i}$ *implies a behavioral standard* $\Delta_i^I(\mathbf{x_i})$ *iff* $\Delta_i^I(\mathbf{x_i}) \in \mathbf{S}(\mathbf{x_i})$ *and* $\Delta_i^I(\mathbf{x_i}) \succeq \Delta^K(\mathbf{x_i}) \quad \forall \quad \Delta^K(\mathbf{x_i}) \in \mathbf{S}(\mathbf{x_i})$.

Note that the same choice can imply different standards for different people: choosing the fair split if the only alternative is to give all to the other person implies fair behavior for an individual who prefers to be fair, but selfish behavior for an individual who prefers to be selfish.[6] [7]

Let $f(\Delta_i^*, \Delta_i^I(\mathbf{x_i}))$ denote the function that defines the cognitive dissonance the individual experiences from her choice. This dissonance depends on how the behavioral standard $\Delta_i^I(\mathbf{x_i})$ that is implied by the individual's choice compares to the individual's preferred behavior Δ_i^*. In particular, behavior that is less preferred than Δ_i^* induces

[5] We are grateful to Joel Sobel for suggesting this kind of exposition. A formal treatment of the emergence of $\mathbf{S}$ is available from the authors upon request.

[6] The implied standard does not have to be unique. If the individual is indifferent between the most preferred standards that are consistent with her choice, this choice implies all of these standards. However, for simplicity and since it does not affect the preference order, in what follows we will treat the set of standards implied by a choice $\mathbf{x}$ as a singleton.

[7] The set $\mathbf{S}(\mathbf{x_i})$ may include some degree of self-deception if self-deception is perfect. Perfect self-deception means that the individual is not aware of the deception anymore, i.e., she truly believes that a certain behavior is consistent with a certain standard. This reflects situations in real life where we are unable to distinguish between "true" meanings we ascribe to a certain behavior and "self-deception" meanings, since we are perfectly convinced by our self-deception.

dissonance:[8]

$$\text{If} \quad \Delta_i^I(\mathbf{x_i}) \prec \Delta_i^* \qquad \Rightarrow \qquad f(\Delta_i^*, \Delta_i^I(\mathbf{x_i})) > 0;$$

In addition,

$$\text{if} \qquad \Delta_i^* \succ \Delta_i^I(\mathbf{x_1}) \succ \Delta_i^I(\mathbf{x_2}) \qquad \Rightarrow \qquad f(\Delta_i^*, \Delta_i^I(\mathbf{x_1})) < f(\Delta_i^*, \Delta_i^I(\mathbf{x_2})) \quad . \tag{1}$$

Cognitive dissonance decreases if the standard implied by a choice is closer to the preferred behavior.

The utility that the individual derives from her choice is then defined as

$$U(\mathbf{x_i}) \quad = \quad u(x_i) - f(\Delta_i^*, \Delta_i^I(\mathbf{x_i})) \quad , \tag{2}$$

with $u(x_i)$ as the standard (monetary) utility that the individual derives from the payoff x_i she herself receives as the consequence of her decision. The general form of this utility function is similar to the models of Rabin (1994) and Konow (2000) but does not explicitly include costs of self-deception.

The individual then chooses $\mathbf{x_i^*}$ to maximize her utility:

$$\mathbf{x_i}^* = max_{\mathbf{x}}\{u(x) - f(\Delta_i^*, \Delta_i^I(\mathbf{x}))\} \tag{3}$$

From (3) it results immediately that if choosing the maximum payoff for oneself in $\mathbf{X}$ implies the preferred behavioral standard Δ_i^*, this payoff is chosen and no dissonance felt independently of the form of the dissonance function. This applies, e.g., to people who consider selfishness as appropriate behavior. In contrast, if choosing the maximum payoff implies $\Delta_i^I(\mathbf{x_i}) \prec \Delta_i^*$, any choice involves either a loss in monetary utility or cognitive dissonance, or both.

Proposition 1 *Let* $f_1(\Delta_i^*, \Delta) > f_2(\Delta_i^*, \Delta) \quad \forall \Delta \prec \Delta_i^*$.

The standard $\Delta_i^I(\mathbf{x_i^*})$ *implied by the individual's optimal choice is (weakly) closer to her preferred standard* Δ_i^* *if her dissonance function is* f_1 *rather than* f_2.

[8] In general, behavior that is "more preferred" than Δ_i^* could induce pride. However, since we model cognitive dissonance here, we abstract from such effects.

The steeper an individual's dissonance function $f(\cdot)$, the more does the dissonance induced by an implied standard affect utility. The optimization (3) then implies that the payoff utility $u(x)$ of choices that do not comply with the individual's self-image (e.g., unfair choices when the individual prefers to be fair) has to be larger for an individual with a steeper dissonance function in order to compensate for this stronger dissonance. The monetary advantages of choices are therefore more likely forgone in favor of a choice closer to the one implied by the individual's preferred standard of behavior. In short, subjects who perceive cognitive dissonance as more unpleasant act less against their standards.

Proposition 2 *Consider a choice set* $\mathbf{X}$ *and two standards* $\Delta_1 \succ \Delta_2$. *Let further* $\mathbf{x_1^*}$ *denote the individual's optimal choice for* $\Delta_1 = \Delta_i^*$ *and* $\mathbf{x_2^*}$ *her optimal choice for* $\Delta_2 = \Delta_i^*$. *Then* $\Delta_i^I(\mathbf{x_1^*}) \succeq \Delta_i^I(\mathbf{x_2^*})$.

Proposition 2 follows from the definitions of the dissonance function in (1) and the utility function in (2). Since a larger deviation from the preferred standard leads to stronger feelings of cognitive dissonance, $f(\Delta_i^*, \Delta_1) > f(\Delta_i^*, \Delta_2)$ for $\Delta_i^* \succ \Delta_2 \succ \Delta_1$, a higher standard of preferred behavior induces stronger dissonance for a given implied standard: $f(\Delta_1^*, \Delta_i^I(\mathbf{x_i})) > f(\Delta_2^*, \Delta_i^I(\mathbf{x_i}))$ for $\Delta_1^* \succ \Delta_2^* \succ \Delta_i^I(\mathbf{x_i})$. According to the optimization in (3), a higher preferred standard then induces behavior that implies a standard which is also (weakly) higher, since the monetary utility has to compensate for a stronger cognitive dissonance. This is independent of the dimension in which a standard is "high": a more social standard implies (weakly) more social behavior, a more selfish standard implies (weakly) more selfish behavior etc.

2.4 Model Extensions

The model can be extended to account for several facts observed in the literature. Including a non-monetary dimension into the choice vector $\mathbf{x}$, the model can account for situations where the same monetary allocation (x_i, x_{-i}) may imply different standards. For example, an unequal allocation may be perceived as unfair if it is the result of an arbitrary choice, but as fair if it results from a real effort task. Similarly, different monetary allocations may imply the same behavioral standard if they differ in the non-

monetary dimension. If the standard implied by a certain choice depends on the non-monetary features of this choice, so does the cognitive dissonance that is induced by it. This dissonance in turn influences the utility an individual derives from a choice, and hence affects her optimal choice. For example, the individual's optimal sharing rule may differ depending on whether she has to share an assigned endowment or a gain of equal size from a real effort task.

Similarly, since the choice set $\mathbf{X}$ is included in the mapping function, the model can explicitly account for the effect of the availability of choices. For example, if only unfair choices are available, choosing the least unfair one must be expected to imply a weakly fairer standard than if fair options are also available. Similarly, it implies a weakly fairer standard if a fair option is chosen when unfair options are available than when only fair options are available.

Finally, by including others' behavior in the mapping function the model can account for the effects of social norms often found in the literature. Consider an individual who prefers to be fair. If she expects or observes others to make less pro-social choices, she may perceive her own choices as implying a fairer standard than if she expects or observes others to make more pro-social choices. In other words, the better others behave, the better the individual herself has to behave in order to comply with her standards. In turn, the better she will behave in order to avoid cognitive dissonance.

2.5 Discussion of the related literature

The main contribution of our model when compared to other *cognitive dissonance-based models* like Konow (2000) and Rabin (1993) is the introduction of different behavioral standards that individuals can apply to their behavior, arising from the various self-concepts people maintain. This allows the model to predict that people can be content with widely varying choices, although they neither deceive themselves nor are ignoring their principles. For example, in allocation situations some people will behave fairly, perceive their behavior as fair and enjoy being in line with their standard, while others behave selfishly, perceive their behavior as egoistic and still enjoy the consistency with their standard. This is not to say that self-deception may not be present. But the model allows for people to be truly content with egoistic choices, without ever incurring costs

of self-deception or cognitive dissonance.

If, independently of their choice, all people were found to apply the same standard to their behavior, this would provide evidence against the model.[9]

The dual self approach (e.g. Bodner and Prelec, 2003; Bénabou and Tirole, 2006; Bénabou and Tirole, 2011) uses a slightly different way to account for self-image as a motivation. It interprets the observer in a social-signaling game (see, for instance, Bernheim, 1994; Bénabou and Tirole, 2006) as the dual-self of the decision maker. The self-signaling decision maker then derives utility from his beliefs about his type preferences as well as from monetary outcomes. However, in such *self-signalling models* individuals are uncertain about their true preferences. Only their choices provide them with signals regarding these preferences. In contrast, models of cognitive dissonance like Konow's (2000) and ours assume that people do know their preferences and can experience a conflict between these preferences and their actions, which is impossible by definition in self-signaling models. Hence, a discrepancy between a person's self-image and (her perception of) her actual choice can only be sustained in models based on cognitive dissonance. With self-signaling, a belief update would lead to a convergence of behavior and (perceived) preferences.

In the model of Bénabou and Tirole (2011) only two types of individuals exist (e.g. low or high altruism). Both make a binary decision, based on which they update the belief regarding their preferences (of belonging to either type). To explain our data, one would have to extend the model to the case of several types and several choices, and specify the process of belief updating beyond a simple Bayesian rule. The model of Bodner and Prelec (2003) could in principle be used to explain our data, given an appropriate specification of the interpretation function that determines the diagnostic utility derived from a choice. This specification, however, requires the assumption of a mental process (which Bodner and Prelec (2003) do not define), which may even be related to cognitive dissonance.

[9] In general, the model allows for all people to apply the same standard. However, if this is the empirical result, the extension to several standards is not necessary and a model like Konow (2000) suffices to explain the data.

3 Experiment

3.1 Participants and Procedures

118 participants were recruited among students from various disciplines at the local university using the ORSEE software (Greiner, 2004). It was compulsory to complete the online personality test a week before the scheduled session in order to take part in the lab experiment. In each session gender composition was approximately balanced and subjects took part only in one session. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007) and took, on average, 60 minutes. The average earnings in the experiment have been €13.43 or $16.63 (including a €2.50/$3 show-up fee for the experiment and an additional €5/$6 for the online questionnaire).

Upon arrival at the laboratory subjects were randomly assigned to one of the computer terminals. Each computer terminal is in a cubicle that does not allow communication or visual interaction among the participants. Participants were given time to privately read the instructions and were allowed to ask for clarifications. In order to check the understanding of the instructions subjects were asked to answer some control questions. After all subjects had answered the questions correctly the experiment started.

At the end of the experiment subjects were paid in cash according to their performance. Privacy was warranted during the payment phase.

3.2 Design

The experiment consisted of three parts: a personality test, a variant of the dictator game, and the assessment of beliefs.

The personality test was administered to the subjects a week before the lab experiment through an Internet platform. It was based on the "self-concept inventory" (SKI) by R. von Georgi and D. Beckmann (2004). As the name suggests, this inventory assesses a subject's concept of herself, in particular in the dimensions *I-strength vs. uncertainty*, *attractiveness vs. marginality*, *trust vs. reserve*, *orderliness vs. carelessness* and *assertiveness vs. cooperation*. Each of these dimensions is covered by 8 questions. We slightly adjusted the survey. Questions of the dimension *attractiveness vs. marginality*

were replaced by questions on other-regarding behavior, to assess the subjects' image of themselves in the dimension of interest for us. In contrast to the established dimensions, the reliability of this scale has not been tested in earlier studies. This has to be kept in mind when interpreting the results. The questions that were used for this scale can be found in the appendix.

The first part of the laboratory experiment consisted of a modified dictator game. Dictators received an endowment of 10 EC (experimental currency), which they could allocate between themselves and a randomly chosen receiver with an endowment of 0 EC. Whether a subject acted as dictator or receiver was determined randomly at the end of the experiment, i.e., all subjects made the allocation decisions. Receivers only learned the outcome, not the choice of the dictator, and dictators knew that. The allocation choice was made for four different scenarios, of which subjects were informed in detail in the instructions. In scenario 1, the dictator's transfer was carried out with certainty, i.e., it reached the receiver and was subtracted from the dictator's account with certainty. In scenario 2, the transfer was carried out with 90% probability only. With the remaining 10% probability, the dictator would keep her 10 EC endowment and the receiver would not get anything, independently of the size of the transfer. In scenario 3, the transfer was carried out only with 80% probability, with 20% both players kept their initial endowments. In scenario 4, the computer decided randomly (50%/50%) whether the transfer was carried out with certainty or only with 80% probability. This scenario was added to control for the effect of letting the computer choose the transfer probability, i.e., adding a layer of ignorance.

Before subjects made their decisions for all four scenarios, they were informed that afterwards it "would be decided" which scenario applied. No specific decision mechanism was mentioned. In fact, after subjects had made all decisions, they were shown a screen with their transfers for all scenarios and could choose themselves which scenario they wanted to apply. Hence, they had the chance to decide whether their transfer would reach the receiver with certainty or not.

In the second part of the laboratory experiment we assessed subjects' first order beliefs regarding the socially appropriate and actually chosen transfers and scenario choices. In order to avoid cognitive overload and strategic behavior of the subjects, detailed

instructions for this part were distributed only after subjects were asked to tell us what they believe is the socially appropriate behavior in part 1 (situation and amount). The initial instructions only mentioned the existence of a second part and stated that further instructions will be given on the screen. After reading the instructions for the second part, subjects practiced the procedure in a similar table for an exemplary task. After all subjects had successfully completed this task, the belief assessment commenced. Subjects were given a table with four columns for the four different scenarios and 5 rows for 5 intervals of transfers (0-2, 2-4, 4-6, 6-8, 8-10). They had to distribute a probability mass of 100% across the 20 cells of this table. Beliefs were elicited in an incentive compatible fashion using a quadratic scoring rule.[10]

A random generator determined for each subject which part of the lab experiment was to be paid. If the first part was chosen, it also determined whether the subject acted as dictator or receiver. This procedure was announced in the instructions. In addition, subjects received a payment of €5/$6 for completing the personality test. The visibility of subjects' choices was kept at a minimum in the experiment. Participation was anonymous and the decision/action was not revealed to the recipient (only the outcome which may be determined in a probabilistic way). It seems hard to imagine that image concerns played a role in this environment.

Finally, subjects filled in a post-experimental questionnaire, where we assessed i) how hard it was for them to choose their transfers and the scenario that would apply; ii) their own judgement of their behavior in the dictator game, i.e., whether on a scale from 1 to 5 they found their behavior generous, fair, rational, clever and egoistic, and iii) which transfer and scenario they thought the receiver had expected them to choose. In addition, we asked subjects, how important it is for them to have and comply with certain principles in life, and how unpleasant they find it not to comply with these principles. We also asked for their age and gender.

[10]Belief elicitation requires quite some additional instructions, especially when incentivizing belief statements and even more so when allowing beliefs to be probabilistic (see Artinger et al. (2010) for a survey). Our instructions emphasized that payoffs are highest if estimates are closest to real values. Instructions did not contain the quadratic scoring rule formula. Instead, subjects were referred to after the experiment if they were interested in the precise calculation. The fact that we experimentally enforce belief statements of course does not mean that participants naturally form such beliefs and are guided by them.

The experimental design shows some similarity with Mahé and Muller (2008), where dictators are also uncertain whether their allocation reaches the receiver. However, since their experiment is meant to test for the presence of warm glow motives in giving decisions, the allocated amount is lost for the dictator, no matter whether it reaches the receiver. *Choosing* a situation where the allocation may fail - the crucial decision in our experiment - is therefore not attractive for the dictator.[11]

3.3 Hypotheses

From the model we can derive the following hypotheses regarding the behavior in the experiment.

Hypothesis 1 *Subjects who ultimately choose a scenario more favorable to them than the certain transfer have relatively more difficulty making this decision, leading to longer decision times and higher reported difficulty in making the decision.*

Subjects' allocation decision involves a tradeoff between the monetary utility derived from keeping as much as possible for oneself, and the cognitive dissonance derived from sending less than what Δ^* would imply. If cognitive dissonance dominates monetary utility, the decision is straightforward, leading to the choice implied by Δ^*. If monetary utility dominates cognitive dissonance or Δ^* implies egoistic behavior, the decision is also straightforward, leading to the egoistic choice. If, however, cognitive dissonance and monetary utility are about equally important, the tradeoff between them is close and the decision becomes more difficult. Subjects who send positive amounts when the transfer is certain but ultimately choose an outcome more favorable to them clearly face such a close tradeoff, as it makes them decide more socially initially but then switch to the more egoistic choice. Subjects who do not thwart their initial intentions by switching to a more favorable scenario may also face a tradeoff, but it is not close enough to make them change their mind. Hence, on average they have and report less difficulty with the decision, and need less time for it.

[11] An alternative would be to introduce the possibility for dictators to *pay* for the allocation to reach the receiver with certainty. This would yield a mirror-image design to ours.

Hypothesis 2 *A higher score in the dimension* other-regarding behavior *of the self-concept inventory is related to more other-regarding behavior cet. par., i.e., higher allocations.*

This hypothesis follows from proposition 2. The dimension *other-regarding behavior* of the SKI serves as a proxy for the standard Δ^* individuals consider appropriate. All else equal, a higher standard should lead to more other-regarding behavior, that is, higher allocations.

Hypothesis 3 *Subjects for whom personal principles are important and who find it unpleasant to deviate from them exhibit more other-regarding behavior.*

This hypothesis is derived from proposition 1. The self-reported relevance of personal principles in life serves as a proxy for the relevance of social principles for the behavioral standard Δ^*. The degree to which deviation from these principles is experienced as unpleasant serves as a proxy for the cognitive dissonance a person experiences if she deviates from Δ^*, i.e., for the steepness of the dissonance function f. The higher an individual's standard of behavior, and the more unpleasant she finds a deviation from it, the more she will show pro-social behavior.[12]

Hypothesis 4 *Higher beliefs regarding the behavior of others are related to more other-regarding behavior.*

This hypothesis results from the extension of the model to include the influence of social norms.

If a subject with a standard of fairness believes that others behave more fairly, this is believed to induce her to also behave more fairly. This effect is based on the argument that if a subject believes that others behave fair, she knows that a given behavior induces a less fair standard of behavior than if she believes others behave unfair. This lower

[12]Some individuals may report a high relevance of personal principles and have in mind principles like egoism or selfishness. However, the evidence suggests that the majority of subjects understands "personal principles" as being social principles. Individual deviations from this interpretation are possible but should not affect the aggregate analysis.

standard of behavior would induce strong cognitive dissonance. In order to avoid or reduce this dissonance, she adjusts her behavior to become more other-regarding. As a result, her beliefs regarding the fair behavior of others induce her to also behave more fairly.

Hypothesis 5 *There is a significant correlation between the self-assessment of subjects and their actual behavior.*

This hypothesis provides a robustness check for the subjects' self-assessment of their behavior. An individual's self assessment of her behavior in the allocation situation (generous, fair, rational, clever, egoistic) serves as a proxy for the standard $\Delta(\mathbf{x})$ induced by her behavior. If the self-assessment is realistic rather than self-deceptive, i.e., a proper proxy for $\Delta(\mathbf{x})$, it should be positively related to real behavior.

4 Results

We first describe how we get data for the variables of interest of our model (self-image, beliefs regarding the behavior of others, assessment of own behavior, and the steepness of the cognitive dissonance function). We then proceed to analyze this data.

4.1 Data

Decisions in part 1 can be separated into i) the amounts subjects sent in the respective dictator situations and ii) what they chose when we let them pick which of the four situations should be executed. We use the four amounts sent to assess a subject's degree of other-regarding behavior in a dictator game. The choice which situation should apply gives us an indication to what extent the subject is willing to yield to the temptation of possibly keeping all the money.

Given the design there are a few possibilities to assess how other-regarding a subject behaves in the dictator situation. One could only look at the first decision without uncertainty or the actual amount sent taking the possible chance of a failed transfer into account. However, subjects provide us with four decisions and it is straightforward

to consider the expected values. That's why we decided to base the degree of other-regarding behavior on the average expected values of the four decisions. Figure 1 shows the distribution.

[Figure 1 about here]

It seems reasonable to assume that values less than 2.5 should be considered as a pro-self choice, although in further analysis we check the robustness of results when we slightly shift this threshold. There may well be a difference in behavior between pro-selfs who give nothing at all and those who give a bit. Hence, we further distinguish pro-selfs into two types: VeryProSelfs with an average expected value of zero and moderate pro-selfs with an average expected value greater than zero, but smaller than 2.5.

We use our second criteria (what they choose when we let them pick which of the four situations should be executed) in order to distinguish pro-socials (average expected value greater or equal than 2.5). If subjects selected the same amount in all four situations and they selected the certain transfer we categorize them as GenuineProSocial. If they picked a situation where the transfer could fail – having entered the same amount in all situations – they are regarded as ShakyProSocials. Some subjects did increase the amount to be sent when the chance of a failed transfer rose. They may have tried to keep the amount to be actually sent at the same level. When they did this in a consistent way,[13] the choice of the situation does not matter. This applies to nine subjects and they are as well regarded as GenuineProSocials. We apply the same logic to further distinguish moderate pro-selfs into genuine and shaky ones. Choice of the certain transfer is not a clear indicator of honesty as for this range of behavior (average expected value greater than zero, but smaller than 2.5) it may well be zero. Hence, when a transfer of zero had been selected the subject was automatically categorized as ShakyProSelf no matter what p was. When a subject sent the same amount and chose the certain transfer or leveled and actually sent a positive amount (s)he was categorized as GenuineProSelf.

[13] We did not check whether their choices comply with expected utility theory as we cannot assume subjects calculate this properly. The sequence of choices needed to look like they tried to level the actual amount. This means strongly monotonically increasing amounts sent or monotonically increasing amounts sent in combination with a final choice to their disadvantage.

This categorization produces 21 VeryProSelfs, 34 ShakyProSelfs, 10 GenuineProSelfs, 26 ShakyProSocials and 27 GenuineProSocials. It is based on a threshold value of 2.5 to distinguish between pro-self and pro-social, and a value of zero to separate VeryProSelfs and moderate pro-selfs.

Table 1: Categorization of the 118 subjects

		Average expected value x of the 4 decisions		
		$x = 0$	$0 < x < 2.5$	$x \geqq 2.5$
		21 VeryProSelfs		
Picked ...	... a situation with a possibly favorable outcome	N/A	33 ShakyProSelfs	26 ShakyProSocials
Resisted to pick ...		N/A	11 GenuineProSelfs	27 GenuineProSocials

We used data from the self-concept inventory (SKI) to measure a participant's self-image with respect to her other-regarding behavior. While the SKI is generally regarded as a very reliable test (reliabilities of the scales is at least .73; see von Georgi and Beckmann, 2004), our six questions on other-regarding behavior were asked for the first time in this framework. Internal reliability is very good (Cronbach's $\alpha = 0.74$), though. The variable *SelfImage* is an individual's average score in the six questions. A high value means a stronger tendency of the individual to have an other-regarding self-image. The questionnaire was conducted anonymously via an online platform, a week before the actual experiment.

In part 2 of the experiment we tried to find out about subjects' beliefs regarding the behavior of others (with respect to the decision they just took in part 1). The literature on social norms is not conclusive about what makes people obey a norm. Bicchieri (2006) distinguishes between empirical expectations (what we expect others to do) and normative expectations (what we believe others think we ought to do). We wanted to elicit both kinds of expectations. Subjects were first asked what they believe is the socially appropriate behavior in part 1 (situation and amount). This is the expected value of the amount they indicated, i. e. the amount multiplied by the degree of certainty of the respective situation $(1, 0.9, 0.8, 0.9)$. The variable *NormExp* expresses subjects' first order beliefs regarding the socially appropriate transfers and scenario choices. The

variable *EmpExp* expresses subjects' first order beliefs regarding the actually chosen transfers and scenario choices. See Figure 2 for the distributions of these variables. It follows that the socially appropriate behavior is not elicited in an incentive-compatible fashion, but *NormExp* and *EmpExp* are.

[Figure 2 about here]

Following Matthey and Regner (2011) we are again interested in the experienced *conflict* and the *decision time* as proxies for the tradeoff between monetary utility and cognitive dissonance. Hence, we took the decision time provided by z-tree when subjects had to select one of the four situations in part 1. Average decision times for the types are 8.05 (VeryProSelfs), 15.03 (ShakyProSelfs), 15.45 (GenuineProSelfs), 20 (ShakyProSocials), and 11.77 (GenuineProSocials). The distributions, see figure 3, are slightly right-skewed. Hence, we use the logarithm of the decision times in the further analysis. In the post-experimental questionnaire we asked subjects how hard it was for them to pick one of the four situations to be implemented.[14]

[Figure 3 about here]

Subjects were also asked to self-assess their behavior in the experiment, that is, whether on a scale from 1 to 5 they found their behavior generous, fair, rational, clever and egoistic. The scores of the last two items were reversed in order to make the scales comparable. Answers for the five items were highly correlated (Cronbach's $\alpha = 0.73$) and we use their average as the variable *SelfAssessment*.

In addition we asked them how important it is for them to have and comply with certain principles in life (*Principles*), and how unpleasant they find it not to comply with these principles (*Dissonance*). We also asked for some background information (age, gender).

4.2 Analysis

First, our interest is whether our results are in line with Matthey and Regner (2011), that is, whether we are able to distinguish subjects who should face a close tradeoff

[14]The precise (translated) text of this question is: "Did you find it easy to decide, which of the four situations in part 1 should take place?"

between cognitive dissonance and monetary utility based on the variables *conflict* and the *decision time* (Hypothesis 1). Types of our categorization that should have two inconsistent psychological cognitions are the ShakyProSelfs and the ShakyProSocials. Both sent positive amounts in the dictator situations. Then they decided to pick a situation with a possibly favorable outcome for themselves – in contrast to the genuine types. The GenuineProSelfs and GenuineProSocials were as well tempted, but resisted picking a situation favorable to them. In other words, they were convinced their original transfer decision (for $p = 0$) is the right one, just like the VeryProSelfs. Hence, we run a multinomial logit regression with robust standard errors, see Table 2. The base outcome of regression I is the joint types VeryProSelf, GenuineProSelf and GenuineProSocial, i.e. individuals whose decision indicates that original and actual choice are equivalent and there is no reason for cognitive dissonance. While VeryProSelfs and GenuineProSelfs/GenuineProSocials clearly differ in behavior, they should still be similar with respect to the variables *conflict* and *decision time*. Regression II takes two aspects into account that may bias the results. First, VeryProSelfs may not only be so convinced about their decision, but also influenced by the fact that they do not really have a choice to make. When zero is to be sent, the probability of the transfer to actually take place does not matter (13 out of 21 nevertheless picked the original transfer). Second, also subjects who leveled (those who equalized the expected amount sent across situations) did not really have a choice to make. Therefore, we exclude the VeryProSelf types and the levelers among the types GenuineProSelf and GenuineProSocial in order to focus on subjects who were in a comparable context. This reduces observations to 82 for regression II.

Both regressions show a positive and significant effect of experienced *conflict* and *decision time* (both at the 1%-level) for ShakyProSocials. There seems to be a strong indication that ShakyProSocials experience a tradeoff between cognitive dissonance and monetary utility. For ShakyProSelfs experienced *conflict* is significant at the 5%-level, the *decision time* is not significant. When the amount to be sent is fairly small (the case of moderate pro-selfs), the consequences of deviating from the originally selected amount (ShakyProSelfs pick $p > 0$) may not create sufficient implications (a 10% probability change of a 1 ECU transfer) to have a noticeable effect. Hence, hypothesis 1 is supported for pro-socials and partly supported for moderate pro-selfs. These results

Table 2: Multinomial logit with type as dependent variable

	ShakyProSelfs		ShakyProSocials	
	I	II	I	II
ExperiencedConflict	.5845	1.154	1.468	1.974
	(.3191) *	(.363) **	(.3776) ***	(.5713) ***
log(DecisionTime)	.1959	.214	.7751	.8296
	(.2044)	(.2796)	(.2296) ***	(.2941) ***
age	-.0655	-.0431	.1182	.1513
	(.1002)	(.1407)	(.1239)	(.1706)
female	.1026	.4334	-.6263	-.0067
	(.4939)	(.6563)	(.6408)	(.8187)
constant	-.6711	-1.283	-8.12	-9.101
	(2.253)	(2.317)	(3.228) **	(4.592)**
N	I: 118		II: 81	
Log pseudolikelihood	I: -104.98		II: -70.84	
Pseudo R2	I: .14		II: .19	

Base outcomes are the joint types VeryProSelf, GenuineProSelf and GenuineProSocial

Significance levels: $***=1\%, **=5\%, *=10\%$

confirm the findings in Matthey and Regner (2011).

We proceed to test specific aspects of the model based on our categorization of subjects into the types VeryProSelfs, ShakyProSelfs, GenuineProSelfs, ShakyProSocials and GenuineProSocials. We estimate an ordered probit model with type as the dependent variable (increasing from VeryProSelf to GenuineProSocial) and robust standard errors, see Table 3.

Generally, results are very much in line with the prediction of our theoretical model. The coefficient of *SelfImage* – an individual's tendency to have an other-regarding self-image – is positive and highly significant. This means hypothesis 2 cannot be rejected. The coefficient of *dissonance* is positive and highly significant, the one for *principles* is significant at the 7%-level. Hence, hypothesis 3 cannot be rejected. Their interaction

Table 3: Ordered probit with type as dependent variable

	coeff.	st.error
SelfImage	.0448	.0165 ***
Dissonance	1.368	.5305 ***
Principles	.7915	.4414 *
Diss * Princ	-.3004	.1286 **
EmpExp	.3099	.0769 ***
age	-.0097	.0409
female	-.1173	.2314
$N = 118$, Pseudo R2: 0.08		
Log pseudolikelihood: -168.94		

Significance levels: $***=1\%, **=5\%, *=10\%$

term appears to have an antagonistic effect (its coefficient is smaller than the main effects). A possible explanation is that if either *dissonance* or *principles* are very high, the other variable becomes less important. The choices of subjects to whom principles are very important do not depend on high dissonance, because these subjects do not violate these important principles. Similarly, if a violation of principles hurts a subject a lot, she does not violate these principles even if they are only moderately important to her. Beliefs about the behavior of relevant others (*EmpExp*) have a positive and highly significant effect.[15] Hence, hypothesis 4 cannot be rejected.

Control variables (age, gender) are not statistically significant. The results are robust for other model specifications (a threshold value of 3 instead of 2.5 to distinguish between pro-self and pro-social and/or a value of 0.5 instead of zero to separate VeryProSelfs and moderate pro-selfs, and a continuous measure instead of the types).

[15] As explained we asked for both normative and empirical expectations in our design. Only the latter have a significant impact on decisions, in line with the findings of Bicchieri and Xiao (2009) who also analysed dictator game settings.

4.3 Discussion

Our model differs from Konow (2000) in two major aspects. First, we allow individuals to have different behavioral standards, i.e., to acknowledge that it is desirable to behave generously, fairly, egoistically, etc. Second, we waive self-deception as an explicit model component. This means that self-deception is assumed to be either perfect, in which case subjects truly believe that they behave according to their standard, or to be imperfect, in which case subjects are aware that they violate their standard. The results support both assumptions. First, the proxy for subjects' behavioral standards, *SelfImage*, is highly significant, implying that different types of subjects indeed have different standards that they find appropriate and apply to themselves. Second, subjects' *SelfAssessment* is highly correlated with their actual behavior in the dictator decisions (Pearson's $r = .77$). We cannot reject hypothesis 5. This implies that – to a substantial extent – subjects indeed assess their own behavior realistically, that is, subjects who allocate nothing or very little are aware that their behavior is egoistic. Mostly they do not deceive themselves to believe that their behavior is fair.

Following up on the dual self models (e.g. Bodner and Prelec, 2003; Bénabou and Tirole, 2006; Bénabou and Tirole, 2011), Grossman (2010b) develops a self-/social-signaling model for a binary dictator game (7,1 vs. 5,5). His 2 x 3 design varies the implementation probability of the dictator choice (1 or 1/3 (with 2/3 the computer randomizes)) and the recipient's information (choice and probability are observable; observer knows the probability and the outcome (not the choice); observer does neither know probability nor observes the choice (only the outcome)). This allows to compare the joint, relative, and independent effects of his self- and social-signaling model in a binary choice situation. When the observer is not informed the model's social-signaling concerns do not predict an effect of a change in the implementation probability. Hence, this condition is used to test the independent effect of self-signaling concerns. Bayesian belief updating is used and it is derived that a self-signaling dictator whose choice counts with 1/3 is more likely to choose (5, 5) than one who knows the choice is implemented for sure. While the experimental results are in line with social-signaling concerns, the predicted effect of self-signaling concerns is not found. Grossman (2010b) concludes that the influence of self-image concerns "... may involve reasoning and cognitive processes not

consistent with a Bayesian signaling model." The results of our experiment suggest that this may in fact be the case. We allow for individual heterogeneity in the moral standard that people maintain, and this individual standard tends to affect the generosity of our participants. Deviating from the moral standard would result in cognitive dissonance. If this psychological cost is too high, participants would stick to the choice their standard implies, while if not, they gain some monetary payoff and experience a conflict.

5 Conclusion

The paper contributes to a growing literature that aims at a better understanding of the nature of social preferences. In general, it points out the importance of factors more subtle than physical outcomes as drivers of this behavior. In particular, it brings attention to self-related aspects as drivers of behavior.

Economic models based on cognitive dissonance can serve as an alternative approach that considers self-image concerns to explain other-regarding behavior. In contrast to outcome-based models of social preferences they are consistent with recent empirical evidence.[16] They explain behavior as being driven by individuals' desire to avoid a divergence between the behavior they consider appropriate for the situation in question and the behavior they actually choose, since such a divergence would lead to two psychologically inconsistent cognitions and induce unpleasant cognitive dissonance.

Our model differs from the seminal work of Konow (2000) in two aspects. First, we lift the assumption of a one-for-all behavioral standard as individual heterogeneity in behavioral standards appears to be a plausible assumption, which is supported by the experimental results of, e.g., Konow (2001). Second, we do not model self-deception explicitly. It is rather an implicit part of our model contained in the mapping function of standards into behavior. Our results seem to suggest that a majority of subjects correctly self-assesses their behavior.

The model's core mechanism is the comparison of a situation's possible choices to one's

[16] See, e.g., the studies of Konow (2000), Dana et al. (2006), Lazear et al. (2006), Dana et al. (2007), Broberg et al. (2007), Larson and Capra (2009), DellaVigna et al. (2009), Grossman(2010a), Matthey and Regner (2011).

behavioral standard for that situation. Any deviating action from that behavioral standard causes subjects to experience two psychologically inconsistent cognitions (their behavioral standard and the deviant behavior). GenuineProSocials, for instance, would follow their behavioral standard in the initial decisions and under uncertainty they would resist the temptation of possibly keeping all the money. Thus, they would not experience cognitive dissonance. For them deviating would be too psychologically costly. In contrast, ShakyProSocials initially would behave according to their behavioral standard, but under uncertainty they would yield to the temptation of possibly keeping all the money. They would experience cognitive dissonance, but the monetary gain outweighs that psychological cost. We derive the following determinants of behavior from the model: an individual's self-image with respect to other-regarding behavior, the sensitivity to cognitive dissonance (the steepness of the dissonance function), and beliefs regarding the behavior of others in the situation in question.

The model predicts i) that a closer tradeoff between monetary utility and cognitive dissonance is experienced by subjects whose choices indicate an inconsistency and ii) a significant relationship between the model variables and the degree of other-regarding behavior. Our experiment, designed to test these specific model components, confirms the model's predictions. We find that subjects whose final decisions deviate from their initial intentions have more difficulty making these decisions than subjects who act in line with their original intentions. Further, the results suggest that subjects with a more other-regarding self-image, subjects with higher social norms for the situation in question, and subjects for whom it is more important to comply with personal principles (and unpleasant not to comply with them) show more other-regarding behavior. This leads to the interpretation that self-image concerns – more precisely the avoidance of cognitive dissonance – play a significant role as driver of pro-social behavior.

The strand of research around other-regarding behavior shows that people are not only driven by their own monetary incentives. Instead, it seems they are motivated to a substantial degree by self- and social-image concerns. From a policy perspective it appears fruitful to provide people with credible information about the consequences of their choices. In particular, consumers – once informed not only about the economic but also the social and environmental consequences of their action – may take pro-social

decisions more often. Having credible information about externalities readily available makes it less likely these consequences would be ignored, and more likely they will be taken into account to comply with one's self-image.

Self-image as a target for policy intervention may also have a more stable effect than monetary incentives. Ariely et al. (2009) analyze the effects of extrinsic motivation (an additional monetary incentive) and social reputation (public visibility of the task) in a donation context. They find a crowding-out of social-image concerns by monetary incentives. In a public setting the level of donations is not increased by introducing additional monetary incentives, while in a private setting monetary incentives raise the donation level. As Ariely et al. (2009) note this has important policy implications with respect to extrinsic motivation. The effect of tax benefits to facilitate adoption of a new environmentally friendly technology should be less successful when they concern a highly visible product (like a hybrid car), since the signalling value of the social reputation effect is reduced by the monetary incentive. There appears to be no indication that such a crowding-out effect of social reputation by monetary incentives applies to self-image effects as well, at least not at a substantial level. Moreover, visibility of the choice should not have a detrimental effect on the self-image. But clearly more research is needed to improve understanding of the interplay between extrinsic motivation, social-, and self-image concerns in determining human decisions.

References

Akerlof, G. A. and W. T. Dickens (1982). The Economic Consequences of Cognitive Dissonance. *American Economic Review*, **72(3)**: 307–19

Ariely, D., Bracha, A. and S. Meier (2009). Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review*, **99(1)**: 544–555

Andreoni, J. and B. D. Bernheim (2009). Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects. *Econometrica*, **77(5)**: 1607-1636

Aronson, E. (1992). The return of the repressed: Dissonance theory makes a comeback. *Psychological Inquiry*, **3**, pp 301–311

Artinger, F., Exadaktylos, F., Koppel, H. and L. Sääksvuori (2010). Applying Quadratic Scoring Rule in multiple choice settings. *Jena Economic Research Papers*, **Vol. 4**, 2010-021.

Beauvois, J.-L. and R.-V. Joule (1996). A Radical Theory of Dissonance. European Monographs in Social Psychology. *Taylor and Francis*, New York, NY

Bénabou, R. and J. Tirole (2006). Incentives and Pro-Social Behavior. *American Economic Review*, **96(5)**: 1652-1678

Bénabou, R. and J. Tirole (2011). Identity, Morals and Taboos: Beliefs as Assets. *mimeo*

Bernheim, B. D. (1994). A Theor of Conformity. *Journal of Political Economy*, **102(5)**: 842-877

Bicchieri, C. (2006) The Grammar of Society. *Cambridge University Press*, New York, NY

Bicchieri, C. and E. Xiao (2009). Do the Right Thing: But Only if Others Do So. *Journal of Behavioral Decision Making*, **22(2)**: 191-208

Bodner, R. and D. Prelec. (2003). The Diagnostic Value of Actions in a Self-Signaling Model. in: *I. Brocas and J.D. Carrillo, eds., The Psychology of Economic Decisions*, Vol.1 , Oxford University Press, 2003.

Bolton, G. and A. Ockenfels (2000). A theory of equity, reciprocity and competition. *American Economic Review*, **100**, 166-193

Broberg, T., Ellingsen, T., and Johannesson, M. (2007). Is generosity involuntary? *Economics Letters*, **94**, 32–37.

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

Dana, J., Weber, R., and Kuang, J.X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*. 33 (2007) 67-80.

DellaVigna, S., List, J., and Malmendier, U. (2009). Testing for Altruism and Social Pressure in Charitable Giving. *NBER Working paper*. 15629.

Ellingsen, T. and M. Johanesson (2008). Pride and prejudice: The human side of incentive theory. *American Economic Review*, **98(3)**: 990-1008

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

Festinger, L. (1957) A Theory of Cognitive Dissonance. *Row Petersen*, Evanston, IL

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics*, **10**(2), 171–178.

Greiner, B. (2004). An Online Recruitment System for Economic Experiments. *University of Cologne working paper series*

Grossman, Z. (2010a). Strategic Ignorance and the Robustness of Social Preferences. *UC Santa Barbara working paper*

Grossman, Z. (2010b). Self-Signaling Versus Social-Signaling in Giving. *UC Santa Barbara working paper*

Harmon-Jones, E. and J. Mills (1999). Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology. *American Psychological Association*, Washington, D.C.

Konow, J. (2000). Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions. *American Economic Review*, **90**(4), 1072–1091.

Konow, J. (2001). Fair and Square: The Four Sides of Distributive Justice. *Journal of Economic Behavior and Organization*, **46**(2), 137–164.

Konow, J. (2010). Mixed feelings: Theories of and evidence on giving. *Journal of Public Economics*, **94**, 279–297.

Larson, T. and C. M. Capra (2009). Exploiting moral wiggle room: Illusory preference for fairness? A comment. *Judgment and Decision Making*, **4**, 467-474.

Lazear, E., U. Malmendier and R. Weber (2006). Sorting in Experiments with Application to Social Preferences. *NBER working paper*. No. W12041 (2006)

Mahé, T. and L. Muller (2008). Is the Preference for Fair Trade motivated by the Warm-Glow of Giving? *mimeo*.

Matthey, A. and T. Regner (2011). Do I really want to know? A cognitive dissonance-based explanation of other-regarding behavior. *Games*, **2**, 114–135.

Mazar, N., O. Amir and D. Ariely (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, **XLV**, 633–644.

Oxoby, R. (2003). Attitudes and allocations: Status, cognitive dissonance, and the manipulation of preferences. *Journal of Economic Behavior and Organization*, **52**, 365–385.

Oxoby, R. (2004). Cognitive dissonance, status, and growth of the underclass. *Economic Journal*, **114**, 729–749.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, **83**(5), 1281–1302.

Rabin, M. (1994). Cognitive dissonance and social changes. *Journal of Economic Behavior and Organization*, **23**, 177–194.

Rabin, M. (1995). Moral Preferences, Moral Constraints, and Self-Serving Biases. *UC Berkeley working paper*, 95-241

von Georgi, R. and Beckmann, D. (2004). Selbstkonzept-Inventar (SKI). Manual. *Hans Huber*, Bern, Switzerland

6 Figures

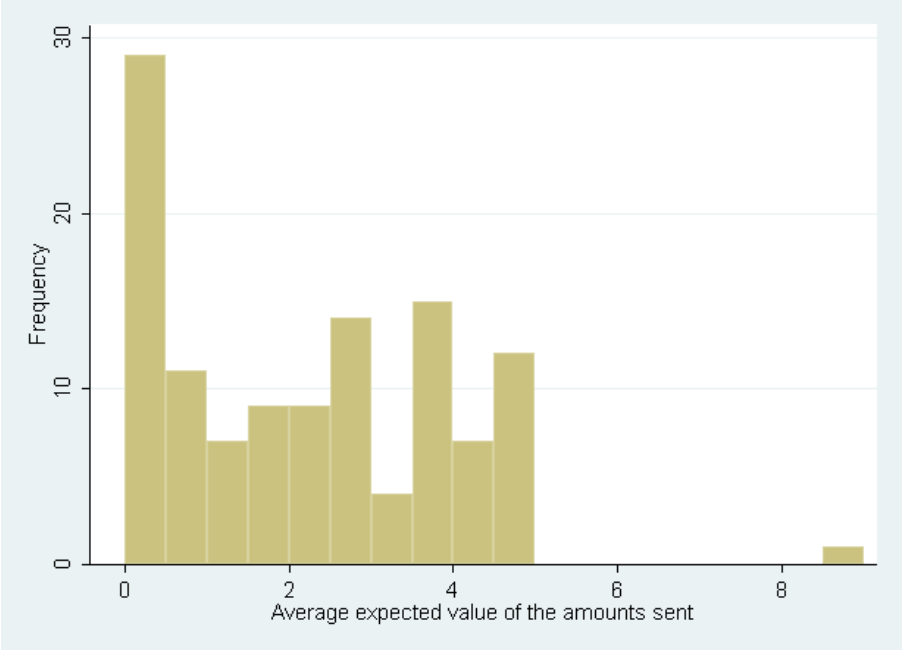
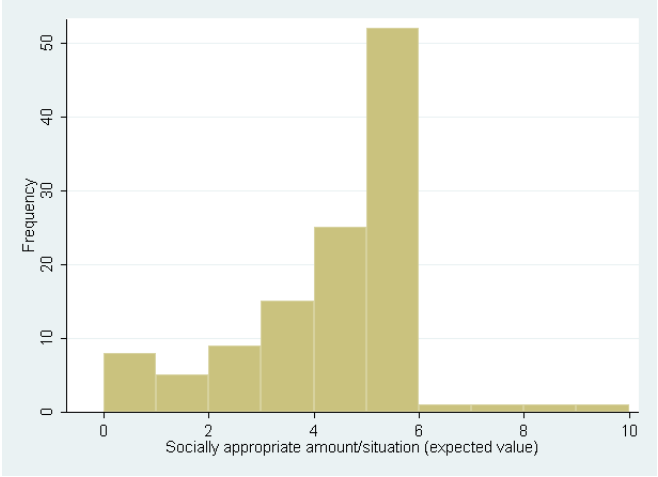
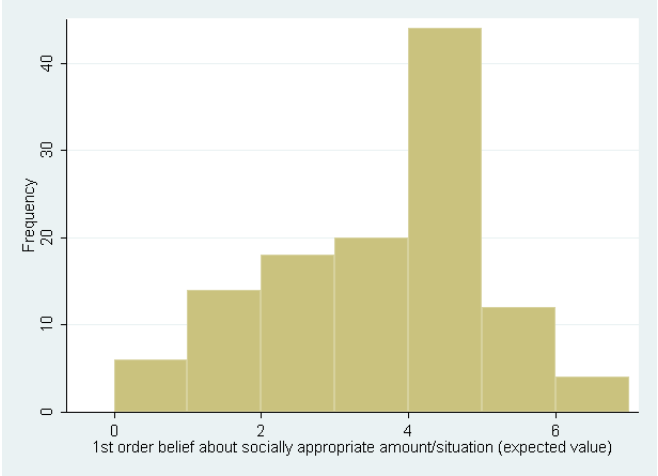


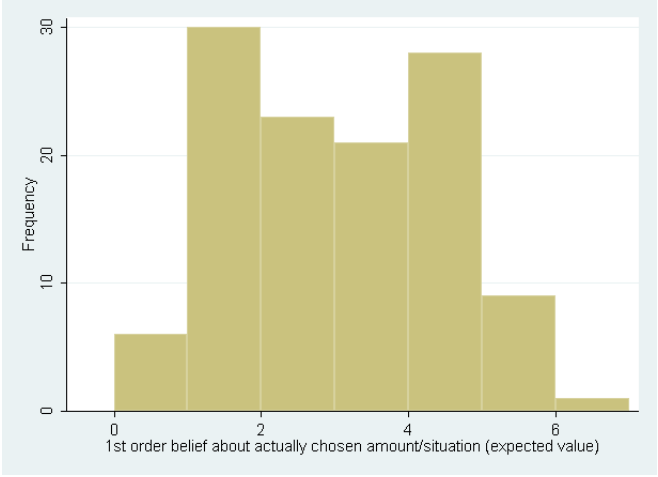
Figure 1: Distribution of the average expected value of the amounts sent



(a) Socially appropriate



(b) 1st order belief about socially appropriate



(c) 1st order belief about actually chosen

Figure 2: Histograms for amounts/situations

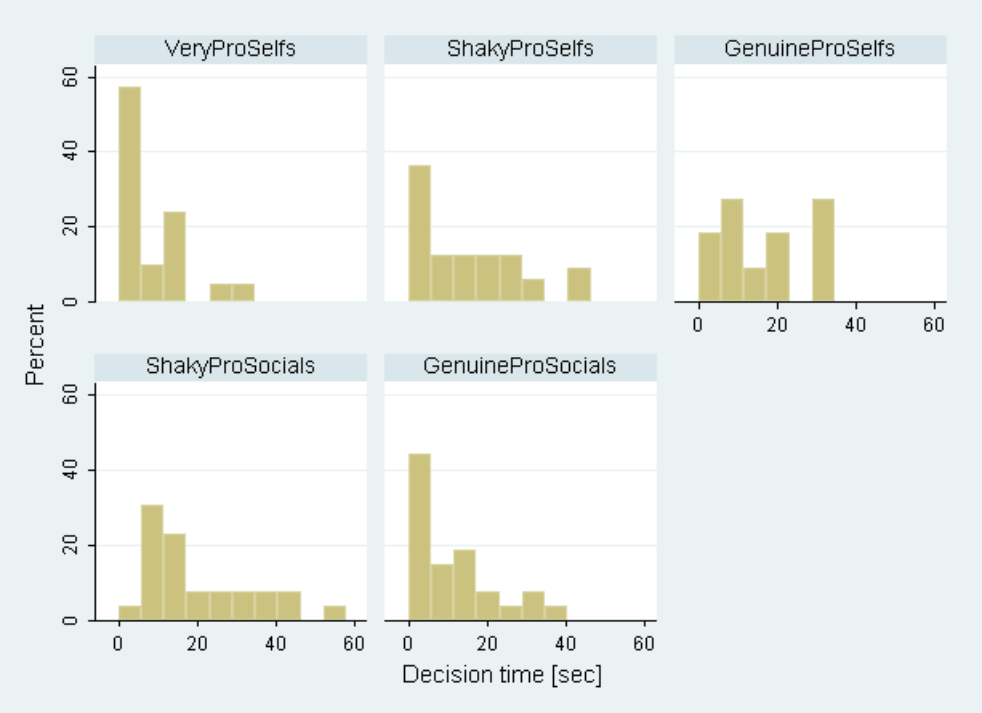


Figure 3: Distribution of decision times by types

Appendix

A. Self-concept Inventory (SKI)

These are the items that were included in the SKI to assess subjects' self-concept in the dimension other-regarding behavior. Items were rated on a 1 to 5 scale between the two extremes. Scores were reversed so that a high value represents a high score in the dimension other-regarding behavior.

a) When making decisions .. I account for the consequences my actions have on others vs. .. I primarily care for my own welfare.

b) Everybody should also consider the interests of others. vs. It is okay if everybody cares mainly about her-/himself.

c) I am rather concerned with my group having plenty vs. with myself having plenty. (or, more generally "faring well"?)

d) I happily share with others. vs. I care primarily for myself.

e) For me it is important that everybody is doing as well as possible vs. that mainly I am doing as well as possible.

f) If somebody is in need, I am happy to share vs. I am not happy to share.

B. Experimental Instructions

Welcome and thanks for participating in this experiment! In this experiment you can earn a certain amount of money, which depends on your and the other participants' decisions in the experiment. **It is, therefore, important that you read the following instructions carefully.** Please note that these instructions are only meant for you and that you are not allowed to exchange any information with the other participants.

Similarly, during the entire experiment it is not allowed to talk to the other participants. If you have any questions, please raise your hand. We will answer your question(s) individually. Please do not ask your question(s) aloud. It is very important that you follow these rules, since otherwise we have to stop the experiment. Please also turn of your mobile phones now.

General procedure

The experiment lasts about 60 minutes. It consists of two parts. In each part you make decisions. Each decision situation will be explained again briefly on the monitor. While you make decisions, the other participants also make decisions which may influence your payoffs.

Your payoff from the experiment depends on your decision and potentially on the decisions of the other participants. But only one of the parts is chosen randomly and you are paid in cash according to the payoff from this part. The exact procedure according to which your payoff is determined is explained below. All amounts in the decision situation are given in EC. They are paid out in exactly the given amount at the end of the experiment. In addition you receive 3 USD /2.50 Euro as a show-up fee. After you filled in a questionnaire the experiment ends and you receive your payoff.

Again an overview of the experiment:

- Reading of the instructions, answering test questions (at the end of the instructions)
- Two parts with decision situations
- Questionnaire
- Payoffs and end of the experiment

Detailed procedure

In each of the first two parts of the experiment you are randomly matched with another participants of the experiment. **You can be sure not to meet the same participants twice.**

Part 1

You get an initial amount of 10 EC. The participant randomly assigned to you for this part gets an endowment of 0 EC. However, you have the opportunity to remit parts of your amount to this participant.

In the process there are 4 different situations:
Situation 1: The remittance will definitely take place. That means that the remitted amount will definitely be withdrawn from your initial amount and added to the amount of the other participant.

Situation 2: The remittance will take place with a probability of 90%. However, with a probability of 10% it will not occur.

- If the remittance will take place, the remitted amount will be withdrawn from your initial amount and added to the initial amount of the other participant.
- If the remittance will not take place, the initial amounts remain as before. 10 EC for you and 0 EC for the other participant, independently of your remittance.

Situation 3: The remittance will occur with a probability of 80%, with a probability of 20% it will not occur.

- If the remittance will take place, the remitted amount will be withdrawn from your initial amount and added to the initial amount of the other participant.
- If the remittance will not take place, the initial amounts remain as before. 10 EC for you and 0 EC for the other participant, independently of your remittance.

Situation 4: The computer decides randomly (50% / 50%), if you are in situation 1 or in situation 2, if your remittance will definitely take place or if it will occur with a probability of 80%.

Course of part 1

In a first step you determine your remittance for all 4 situations. At that point it will be decided which situation will occur. After that the computer determines if the remittance will take place or not (except for situation 1 in which the remittance will definitely take place). Afterwards the respective remittance will be carried out.

Payoff

If at the end of the experiment part 1 is to be paid out, there are two options:

1. You receive your initial amount of 10 EC, if necessary less your remittance.
2. You don't get an initial amount (0 EC), but the amount the participant assigned to you has remitted to you.

Which of both options will occur, will be determined randomly (50% / 50%), if part 1 will be determined as relevant for your payoff. The determination of your payoff will be explained in detail in the course of the instructions.

Part 2

Instructions for this part will be shown on your monitor.

Your payoff from the experiment

For your payoff only **one** of the parts is relevant and it is chosen randomly (step 1). If part 1 is chosen for your payoff, a second step will determine if your own decision is relevant for your payoff or the decision of the decision of the participant who has been been assigned to you for this part (step 2). If you or the participant assigned to you have made several decisions in the chosen part, the decision relevant for your payoff will be chosen randomly (step 3).

Example of the determination of the payoff:
Step 1: Part 1 is chosen as relevant for your payoff.
Step 2: The decision of the participant assigned to you is chosen as relevant for your payoff.
Step 3: As only one of four situations will eventuate, the participant has only made one decision in this part. In particular, he has for example remitted „X“ EC to you in the situation eventuated. Therefore you gain „X“ EC.
You will receive the determined payoff in cash immediately after the ending of the experiment (that means after having completed the final questionnaire). The other participants will not be able to see your payoff.

Explanation concerning probabilities

Probabilities of 80% or 90% can be imagined as following:

You blindly pull one ball out off a bowl with 100 balls. If the bowl contains 20 red and 80 black balls, the probability to pull a red ball is 20%, the probability to pull a black ball is 80%. If the bowl contains 10 red and 90 black balls, the probability to pull a red ball is 10%. The probability to pull a black ball is 90%.