CESifo WORKING PAPERS

8070
2020

January 2020

When a Nudge Backfires: Combining (Im)Plausible Deniability with Social and Economic Incentives to Promote Behavioral Change

Gary Bolton, Eugen Dimant, Ulrich Schmidt

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 8070*

When a Nudge Backfires: Combining (Im)Plausible Deniability with Social and Economic Incentives to Promote Behavioral Change

Abstract

Both theory and recent empirical evidence on nudging suggest that observability of behavior acts as an instrument for promoting (discouraging) pro-social (anti-social) behavior. We connect three streams of literature (nudging, social preferences, and social norms) to investigate the universality of these claims. By employing a series of high-powered laboratory and online studies, we report here on an investigation of the questions of when and in what form backfiring occurs, the mechanism behind the backfiring, and how to mitigate it. We find that inequality aversion moderates the effectiveness of such nudges and that increasing the focus on social norms can counteract the backfiring effects of such behavioral interventions. Our results are informative for those who work on nudging and behavioral change, including scholars, company officials, and policy-makers.

JEL-Codes: C910, D640, D900.

Keywords: anti-social behavior, nudge, pro-social behavior, reputation, social norms.

Gary Bolton
University of Texas at Dallas / USA
gbolton@utdallas.edu

Eugen Dimant
University of Pennsylvania
Philadelphia / PA / USA
edimant@sas.upenn.edu

Ulrich Schmidt
Kiel Institute for the World Economy
Kiel / Germany
Ulrich.Schmidt@ifw-kiel.de

This version: January 17, 2020
The most recent version of the working paper can always be downloaded following this link:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3294375
This work benefited from discussions with Roland Bénabou, Anastasia Danilov, Christine Exley, Ernst Fehr, Gianluca Grimalda, Larry Katz, Judd Kessler, George Loewenstein, Katherine Milkman, Muriel Niederle, Todd Rogers, Al Roth, and Silvia Sonderegger. Helpful input was provided by Jana Freundt and two research assistants. We also thank participants at the 2018 Norms and Behavioral Change Workshop and the 2019 Innsbruck Winter Summit, as well as seminar participants at Heidelberg University, Humboldt University, the Kiel Institute for the World Economy, the Lab@DC Behavioral Research Team, the UPenn Social and Behavioral Science Initiative (SBSI), and Yale University for feedback. We acknowlege financial support from the German Science Foundation (DFG) via the research unit “Design Behavior” (FOR 1371).

1. Introduction

Much of the existing behaviorally informed policies have focused on improving individual and collective well-being, with some approaches having yielded more success than others (Thaler and Sunstein, 2008; Benartzi et al., 2017; Brandon et al., 2017). Importantly, however, much less scholarly attention has been paid to the unintended and sometimes detrimental consequences of nudging (Reijula et al., 2018), which go beyond simply unsuccessful interventions, but may lead to actual backfiring effects that worsen outcomes (for a discussion, see Sunstein, 2017, Gino et al., 2019).[1] A few recent examples of such nudge interventions include an increase in energy consumption (Ayres et al., 2013), increase in prescription of antibiotics (Hallsworth et al., 2016), increase in deviant behavior (Dimant et al., 2019), reduced savings in a 401(k) (Beshears et al., 2015), decrease in organ donor registrations (Behavioral Insights Team, 2015), a decline in sustainable food choices (Richter et al., 2018) and support for environmental policies (Hagmann et al., 2019). Taken together, these findings are insightful because they highlight the potential risks of behavioral interventions affecting societal and economic outcomes on both the individual and collective level (Madrian, 2014; Allcott and Kessler, 2019; Bicchieri and Dimant, 2019).

Importantly, however, scientific research also needs to illuminate the scholarly debate on *why* such backfiring might arise. Here, we investigate why and when nudging can backfire and what to do about it.[2] Our findings connect the backfiring phenomenon to the literature on inequality aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) and the ability to counteract such to the social norms literature, in particular to the recent insights on norm-nudging (Cialdini et al., 1990; Bicchieri, 2006; Bicchieri and Dimant, 2019). To the best of our knowledge the experiments we report here are the first to establish such links.

To answer these important questions, we develop a novel experimental setup and utilize high-powered studies both in the laboratory and online, consisting of a total of 11

[1] Given the methodological difficulty of comparing nudge interventions (Szaszi et al., 2018), which is amplified by inherent publication bias of null-findings in social sciences, it is challenging to credibly estimate how many nudge interventions fail. Anecdotal evidence suggest that up to 50% of the interventions fail, as argued by David Laibson in his talk at the Wharton Business School on November 12, 2018. The number is based on his personal – and likely positively biased – scholarly experience.

[2] We will refer to *backfiring* in situations where the intervention produced a detectable behavioral reaction contrary to what was intended (and subsequently worsening welfare), rather than simply being ineffective (to which we will refer as such).

treatments in excess of 2,000 participants. Our principal contribution is to present the fragility of a particularly prominent nudging approach in which observability of behavior is being used as a catalyst for behavioral change. More specifically, we first examine the robustness of such nudges by systematically varying the degree of observability of one's actions and the degree to which the parties interact with each other. Our working assumption is that observability affects the degree of plausible deniability both directly (through economic incentives following from reputation concerns) and indirectly (through social incentives following from image concerns), which in turn mobilizes behavioral change (e.g., Von Hippel and Trivers, 2011; Rogers et al., 2018; Ali and Bénabou, 2019).

In fact, we find that observability absent some material incentive does not have much value in promoting pro-socialite and is even suggestive of a slightly negative overall impact. We establish the robustness of this unexpected result in follow-up online experiment where we used a larger, more heterogeneous, and generalizable sample of participants to also investigate the potential mechanisms in more detail. These experiments focus on the self and social observation conditions and several variants thereof. For one, the experiment establishes the robustness of the result that mere observation without material incentives can indeed yield a backfiring effect, i.e. significantly more pronounced anti-social behavior. We then follow-up with treatments that (i) investigate the mechanism behind the backfiring effect, along with treatments that (ii) test whether increased focus on the norm can nudge people to greater pro-social behavior when they are being observed.

It turns out that observability induces anti-social behavior such that the overall effect of the nudge becomes negative. We show that one important driver for this backfiring is inequality aversion as in our experiments disadvantageous inequality has stronger effects when one is observed by others. This can be explained by reference group dependence, i.e. a subject becomes more important in the reference group if she can observe ones actions (Hogg and Turner, 1987; Bicchieri, 2006, see also insights from identity economics Akerlof and Kranton, 2000; Kranton, 2016), an insight that has been left rather vague in the original inequality models (see discussions in Fehr and Schmidt (1999, pp. 821-822) and Bolton and Ockenfels (2000, p. 189); for related insights on narrowly framed equity concerns see Fisman et al., 2017; Exley and Kessler, 2018). In addition, the insights advance our understanding of the role of context effects in altruism as well as norm-learning and -following (Krupka and Weber, 2013; Bicchieri et al., 2019a; Dimant, 2019).

The main conclusion that follows from out study is that nudging individuals via social observation per se can have little pro-social benefit in an anonymous setting like ours and

in which both anti-social and pro-social actions are attainable. Rather, in the absence of enforcement mechanisms, the power of social observation rests largely with its ability to nudge decision makers into making the existence and applicability of norms salient.

In what follows: Section 2 discusses the relevant streams of literature and how we position our contribution. Section 3 details the laboratory experiment, develops a theoretical foundation for the testable hypotheses, and presents the results. We introduce the follow-up MTurk experiments and discuss the results in Section 4. Section 5 concludes.

2. Relevant Literature and Mechanisms of Interest

As discussed above, we can think of social observation as a nudge intended to get people to think about the social consequences of their behavior, usually with the aim of encouraging pro-social behavior or reducing anti-social behavior.[3] This has become a particularly popular approach in the current 'nudge revolution' due to its fairly frugal implementation. We follow the definition by Thaler and Sunstein (2008, p. 6) in what constitutes a nudge: "A nudge is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives." Our main intervention of interest – the study of how changes in the observability of one's action – is consistent with this categorization since we hold both the action space of our decision-maker and the monetary rewards constant across treatments.[4]

Historically, some of the most prominent evidence for pro-social behavior has been gathered from the dictator game (Forsythe et al., 1994), in which subjects may voluntarily transfer a substantial fraction of their own endowment to other individuals. From early on, a critical point of debate has been the role of observability as a catalyst for observed generosity. Specifically, whether pro-social behavior is driven by the economic consequences of the reputation that results from being observed, or whether the act of being observed is in itself a sufficient pro-social motivator (Andreoni and Bernheim, 2009).[5]

[3] Here we will equate pro-social behavior with actions that increase the welfare of others and anti-social behavior with actions that decrease the welfare of others. See Dimant (2019) for a related discussion.

[4] Note that our contribution is not limited to the interpretation of representing a nudge intervention, but also falls under the general umbrella of the social preference literature constituting that individuals are sensitive to (non-)monetary incentives (e.g., Akerlof, 1980; Milgrom et al., 1990; Bolton et al., 2004; Andreoni and Bernheim, 2009; Bolton et al., 2013).

[5] Social approval is deeply ingrained in human nature and even when observers cannot reciprocate behavior through a reputation system, social observation might still act to nudge the observee's decision-making towards pro-social action (Akerlof, 1980; Lacetera and Macis, 2010). In the absence of observability, the act

Beyond tit-for-tat reputation mechanisms, the empirical relation between observability and social behavior is less clear and has produced conflicting evidence, which we try to reconcile in our experiment. Studies analyzing the effect of observability on pro-sociality find a variety of results ranging from positive effects (Haley and Fessler, 2005; Lacetera and Macis, 2010; Linardi and McConnell, 2011), to negative effects (Lambarraa and Riener, 2015), to no effect (Dufwenberg and Muren, 2006; Cason et al., 2016; Matland and Murray, 2016), in particular when the observer is an uninvolved third party (Alpízar and Martinsson, 2013; Festré and Garrouste, 2015). In contrast, evidence suggests that there is a strong effect of observability if the observer interacts with the giver in later rounds of the game, i.e. during which (monetary) reputation concerns are present (Bradley et al., 2018). Importantly, however, and in the spirit of our investigation, recent evidence has produced the insight that observation can backfire under certain circumstances; i.e., if it is viewed as coercive. For instance, past charity donors who were reminded by email to donate, were observed to do so at first but then observed to remove themselves from the mailing list. This is evidence to suggest that a reminder, in the form of observation, created resistance on the part of the donors (Damgaard and Gravert, 2018). Research that considers the effects of observability on anti-social behavior finds that observing and being observed can lead to behavioral contagion of deviant behavior and the erosion of norms (Falk and Fischbacher, 2002; Gino et al., 2009; Rilke et al., 2018; Bicchieri et al., 2019a; Dimant, 2019).

In contrast to this literature, we are not investigating how observation of others' behavior affects one's own behavior. Instead, we are interested in answering: how does the exposure of one's actions affect one's decisions to engage in pro- or anti-social behavior? Our aim is to disentangle the mechanisms through which observability of behavior nudges behavior. In particular, we focus on the prominent channels of altruism/self-image, social image, and the reputation mechanism. We systematically vary the degree of observability of one's actions and the degree to which the parties interact with each other.[6]

Typically, experimental analyses of indirect reciprocity involve dictator games where

of giving in a dictator game can result from intrinsic altruism or the desire to maintain a positive self-image (Bénabou and Tirole, 2006; Hamman et al., 2010; Bénabou and Tirole, 2011; Bénabou et al., 2018). One well-established channel in this context is warm glow of giving (Andreoni, 1990).

[6]Strictly speaking, our paper investigates two types of nudge interventions (*observability* in Sections 3 and 4.4.1 as well as *norm focus* in Section 4.4.3), whereas Section 4.4.2 investigates the mechanism of the backfiring beyond what would constitute a nudge by exogenously varying a participant's initial wealth. These additional interventions are important to uncover the exact mechanisms of the nudge interventions.

only a giving option is present (within the context of reputation concerns in charitable giving see, e.g., Kajackaite and Sliwka, 2017). In contrast to this and much of the related nudging literature more generally that addresses the role of observability, we employ a framework where both pro- and anti-social behavior is simultaneously possible. For this purpose, we refine the give-or-take dictator game setting (List, 2007; Bardsley, 2008; Dimant, 2019) and augment the setup with a charity and other (un)involved parties.[7] Thus, our design allows us to study the impact of an observability nudge in a indirect reciprocity setting where anti-social behavior is possible as well (see also Balliet et al. 2014).

It is worthwhile putting a more recent stream of literature in perspective. A central finding in behavioral economics is that people display pro-social behavior, at least under certain institutional conditions, i.e. when compliant (deviant) behavior can be enforced in both monetary and non-monetary terms (Becker, 1968; Xiao and Houser, 2011; Balafoutas et al., 2014; Cooper and Kagel, 2016; Bicchieri et al., 2019c). Observability is also the basis for institutionalized reputation mechanisms, which leverage economic incentives and have been shown to facilitate social welfare enhancing outcomes, such as increasing coordination and trust and enabling trade relations (Buchan et al., 2002; Bolton et al., 2004, 2013). Also, indirect reciprocity (Alexander, 1987; Nowak and Sigmund, 1998a,b; Engelmann and Fischbacher, 2009; Grimalda et al., 2016), a major explanation for the evolution of human cooperation, relies on the reputation gained from observable previous actions. Crucial to the effectiveness of observability in these contexts is the possibility that the observer can reciprocate or enforce the behavior of the observee. Similar to earlier studies, we find that reputation mechanisms that allow for a (monetary) enforcement component increase pro-social behavior and curtail anti-social behavior relative to a self-signaling condition in which the only observer is oneself. This shows that the evidence on reputation mechanisms and indirect reciprocity is robust with respect to a generalization of the experimental design where both pro- and antisocial behavior are possible. However, we also find that social observation, absent the tit-for-tat lever, has little effect on overall pro-social behavior and, if anything, indicates that anti-social behavior can increase.

In addition to examine when and how backfiring occurs, we also shed light on (i) the

[7] A recent debate centers around the question whether giving in the dictator game really displays intrinsic altruism or is either (i) an artifact of the experimental design (List, 2007; Bardsley, 2008) as people can only give but not take and, therefore, their choice set is artificially censored or (ii) caused by the observability of decisions by other persons such that giving is due to social image concerns rather than due to intrinsic altruism (Dana et al., 2006, 2007; Andreoni and Bernheim, 2009; Lazear et al., 2012; Cappelen et al., 2017).

reasons for such backfiring and (ii) ways to counteract it. For (i) we are motivated by the literature on inequality concerns in conjunction with the hypothesis that the focus of such concerns is a function of whether or not one's actions are observed and by whom, e.g. because the reference group to which one compares onself to changes (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000, for a recent and relevant discussion see Fisman et al., 2017; Exley and Kessler, 2018). We present evidence that this is a likely channel for the observed backfiring effect. For (ii), we test the hypothesis that nudging people to think about a descriptive norm of what others do affects the efficacy of the observation intervention as suggested by Cialdini et al. (1990, 1991) and instrumentalized by Krupka and Weber (2009), which turns out highly effective.

3. Laboratory Experiment

While the impact of observability on inducing pro-social behavior can be exerted through various mechanisms such as reputation concerns, shame, feelings of increased accountability, or a change in social norms (e.g., Bohnet and Frey, 1999; Prat, 2005; Bénabou and Tirole, 2006; Ernest-Jones et al., 2011; Ekström, 2012; Bursztyn and Jensen, 2017; Bicchieri et al., 2019a), existing literature reveals mixed results and a distorted picture of when and why certain nudges that capitalize on observation and harness reputation do or do not work. We attempt to shed more light on these conflicting results in a controlled environment by disentangling the several behavioral channels at play, while simultaneously reducing observability to actions alone in order to establish a clear causal chain. We achieve this in three steps: first, we investigate the role of observability in a controlled laboratory experiment in which we systematically vary the observability of one's actions by others as well as the (non-)monetary relationship between the observer and observee. We establish empirically that some interventions achieve the desired effect while others fail. Next, we break down the studied behavioral mechanisms further into its basic elements and establish through various high-powered online experiments that – absent proper economic incentives – nudge interventions that rely on social incentives can go rogue and even backfire. Lastly, a set of similarly powered treatments seek to shed light on how to counteract the observed backfiring by borrowing from the social norms literature (Bicchieri and Dimant, 2019).

For our first step, we employ the following design in the laboratory: A decision maker can donate money from her own account to a charity or can transfer money from the charity's account to her own. In the baseline condition, the only player who can observe

the decision maker's action is the decision maker herself (the self-image treatment, *SelfSig*). We complement the baseline with two observability treatments. In one treatment (social image condition, *SocSig*), an uninvolved third party observes the decisions of the decision maker (dictator). In another condition (reputation treatment, *Reputation*), decisions are observed by a person the dictator interacts with in a later stage of the game. Data from this treatment provides a benchmark, arguably an upper bound, on how much pro-social behavior we might expect from self or social image alone. For steps two and three, we remove the reputation channel in our online experiments so we can focus solely on the role of social incentives (in the form of social signaling concerns). In these steps, we vary both the degree of observability and the payoff inequality among the observer and the observee.

3.1. Design

Our basic design capitalizes on an extended one-shot dictator game with a taking option (most closely mirroring List, 2007; Bardsley, 2008; Dimant, 2019, but see also Coffman, 2011). Participants were randomly assigned to one of three roles (A, B, or C) that differ in their action space:

- **Player A** moves first and plays the role of a dictator who is assigned to a charity. Both player A and the charity start with an initial endowment of 100 ECU (= €10). Player A must decide whether to (i) give some or all of her own endowment to charity, (ii) not change the equal split, or (iii) take some or all of the charity's endowment and add this to her own endowment.[8]

- **Player B** moves second and plays the role of a dictator who is in charge of a cash box that is separate from her own endowment. Both Player B and the cash box have an initial endowment of 100 ECU. The final payoff of B is always equal 100 ECU in that it is unaffected by any of her own or other players' decisions in the experiment. Player B must then decide whether to transfer any or all of the money from the cash box to Player A. Any amount left in the cash box will be returned to the experimenter and can neither be kept nor added to her own endowment. Hence, Player B has no personal monetary incentive for her actions, as she cannot enhance her own payoff

[8] We include a welfare multiplier of 2 for giving decisions, where the experimenter will double every monetary unit transferred to the charity account.

by leaving money in the cash box.[9]

- **Player C** has the role of a passive observer who does not engage in active decision-making. Just as with Player B, Player C has an initial endowment and a final payoff of 100 ECU. Player C never observes what Player B does with the cash box, which is common knowledge among all participants.

A crucial feature of our experimental design is that we systematically vary the observability of the decisions that participants make across treatments.[10] Our experiment consists of one Baseline (*SelfSig*) specification and two treatments. In the SelfSig condition, henceforth referred to as **SelfSig**, Player A's behavior remains unobserved by other participants; any reputation concerns are absent and each decision is merely self-signaling.[11] That is, while Player B knows that Player A is making some decision towards the charity, Player B makes her decision regarding the separate cash box *without* knowing what exactly Player A did. Figure 1 displays the action space and the order of actions for all participants in this SelfSig condition. In **Treatment 1**, henceforth referred to as **Reputation**, Player B observes Player A's behavior, i.e. the amount given to or taken from the charity, before making her own decision about whether to give some or all of the money from the cash box to Player A. This is public knowledge and Player A is made aware of this circumstance prior to making her decision towards the charity. Therefore, both monetary and non-monetary concerns are at play as Player B can condition her cash box decision on Player A's behavior. As explained in the Introduction, this treatment serves as a benchmark to compare non-monetary observation effects. We detail the mechanism in Figure 2. In **Treatment 2**, henceforth referred to as **SocSig**, Player C, rather than Player B, can observe the decision of Player A but has no impact on Player A's payoffs. Just as in the SelfSig treatment,

[9] Again, we include a welfare multiplier so that the experimenter will double any amount given to Player A. This is necessary to give Player A the incentive to be rewarded by Player B for being pro-social towards the charity and leave the experiment with more money than she started with. For example, if Player A decides to give all of her money (100 ECU) to charity, this amount is then doubled and added to the charity account with the initial 100 ECU, making it a total of 300 ECU. If, in return, Player B decides to give all of the money (100 ECU) in the cash box to Player A, this amount is then doubled and Player A ends up with 200 ECU. In consequence, a fully pro-social behavior on the side of both Player A and Player B leads to a substantial welfare increase.

[10] Observability in our experiment is limited to actions - participants were not able to identify each other.

[11] Admittedly, self-signaling exists in all treatments. In this treatment, however, self-signaling is the exclusive channel. In line with the theoretical predictions, the extent of self-signaling is the same across treatments, and what varies are the additional channels such as reputation and social signaling.

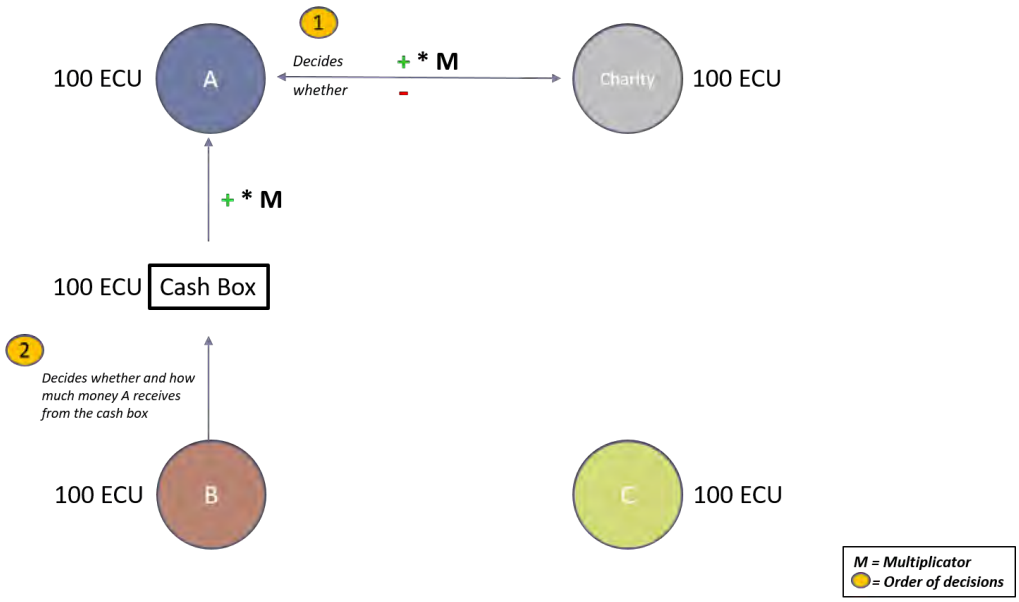

Figure 1: Experiment 1 - Baseline Condition (SelfSig).

Player B makes her cash box decision without being informed about Player A's behavior. Hence, Player A's decision is being observed by a passive player which allows us to separate reputation concerns from social concerns. As was the case before, the setup is public knowledge and every player is made aware of the action space of each other player prior to the first decision being made. In consequence, in addition to self-signaling concerns that are existent in all treatments, Player A's behavior now also bears social signaling value. We present the details of this treatment in Figure 3.

The payoff structure employed in our experiment retains incentive compatibility while accounting for potential crowding out effects caused by strategic considerations. We achieve incentive compatibility by randomly selecting one group at the end of each session and then using the decisions of the group members to determine their own and the charity's payoffs. The payoff of the participants in the remaining n-1 groups is equal to their initial endowment (100 ECU), independent of their decisions throughout the experiment.[12] The amount of the chosen charity payoff is then given randomly to one of the three pre-announced char-

[12]This implementation of a payoff function closely follows Bicchieri et al. (2019a) and Dimant (2019). This incentive scheme is in line with suggestions made by Charness et al. (2016) and retains incentive compatibility as theoretically argued by Azrieli et al. (2018).

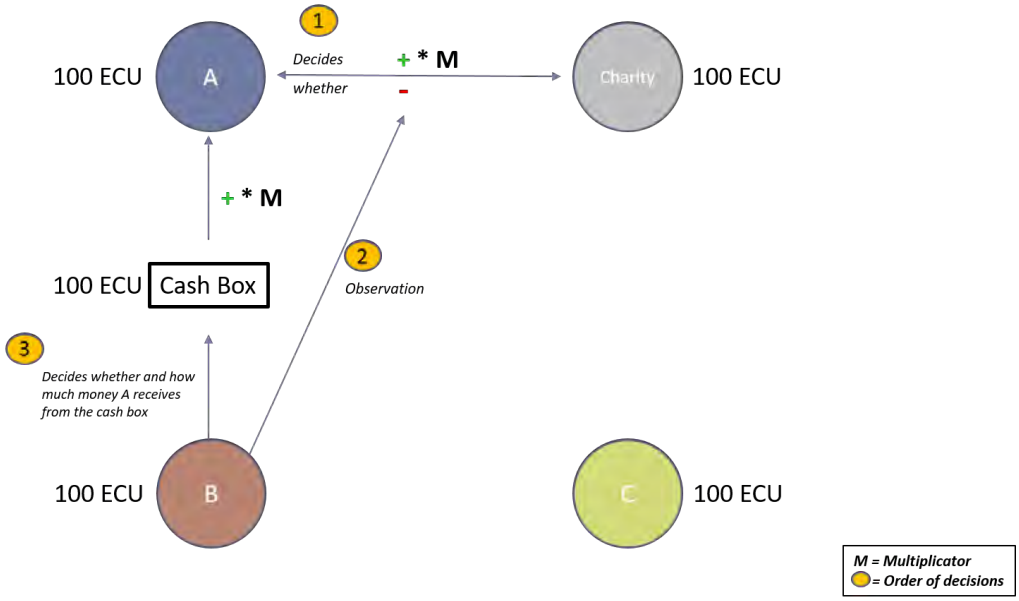

Figure 2: Experiment 1 - Treatment 1 (Reputation).

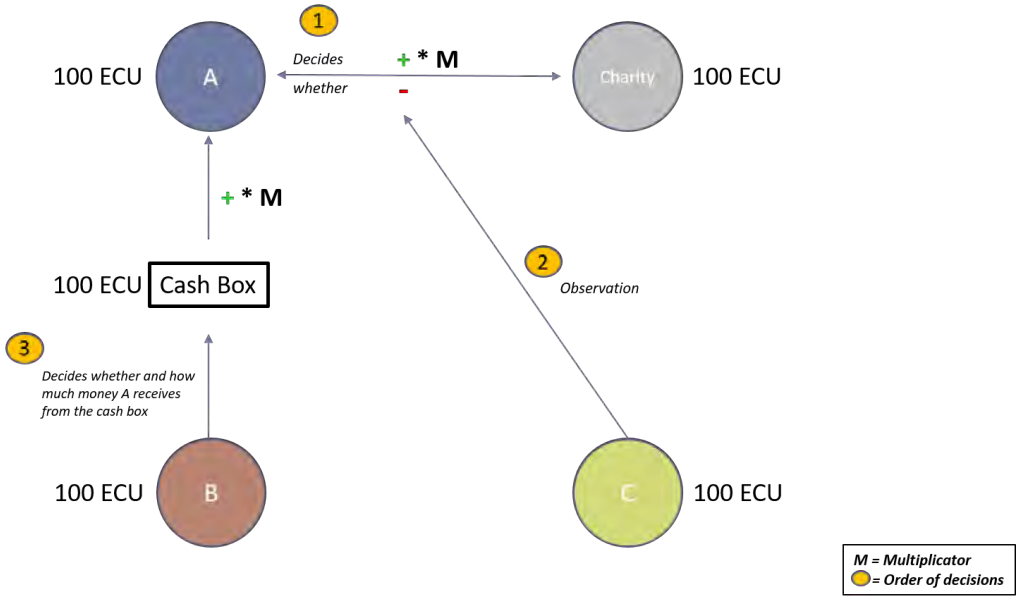

Figure 3: Experiment 1 - Treatment 2 (SocSig).

ities. This procedure is common knowledge among all participants and is explained in the instructions given on the participants' screens.[13]

3.2. Participants and Procedures

For the laboratory experiment, we recruited 423 subjects from various disciplines at the experimental laboratory at the University of Kiel, Germany. Observations were collected in triads, totalling 138 participants for both SelfSig and Reputation each, and 147 participants for SocSig, respectively. Participants were on average 24.6 years old and 54.7% were female. The experiment was programmed and conducted with z-Tree (Fischbacher, 2007). We ran a total of 28 sessions, each session lasting approximately 60 minutes. See Table 1 for details.

Treatments	Endowment (A / Charity / B / C)	Who Observes A?	# Player A
SelfSig	100 / 100 / 100 / 100	Nobody	46
Reputation	100 / 100 / 100 / 100	B	46
SocSig	100 / 100 / 100 / 100	C	49

Table 1: Descriptive Statistics for Player A.

Upon arrival, subjects were randomly assigned to a computer. After the on-boarding process was completed, instructions were distributed and participants received sufficient time to read them carefully and to ask any clarifying questions. Each participant received the same basic instructions regardless of her role or treatment. Treatment variations were announced on the screen of the participants' computers. Once the experiment started, participants had to answer a series of comprehension questions to ensure the instructions were clearly understood. After each participant answered the questions correctly, the experiment continued with the random assignment of participants to roles. Depending on their assigned role, participants had to make decisions that simultaneously affected their own payoff and that of a charity, or that only affected the payoff of another participant, as indicated above. At the end of the experiment subjects filled out a questionnaire and were privately paid in cash according to the outcome of the experiment; the donation decision of one randomly selected participant was implemented.

[13] More generally, all treatment variations were always common knowledge in all treatments since each participant, regardless of her role, received the complete set of instructions.

3.3. Theory and Behavioral Predictions

Let us denote the money Player A transfers to the charity by x where negative values of x indicate taking money from the charity. Due to Player A's initial balance, we have $-100 \leq x \leq 100$. Thus, $100 - x$ is the money Player A keeps in her private account. Additionally, Player A may receive money which Player B assigns to her from the cash box. We denote this transfer by y. We assume that the behavior of A is motivated by a combination of the following factors:

1. Self-interest, i.e. the money kept from the initial endowment, the money taken from the charity (in total 100 – x) and the expectations about the money Player B will transfer from the cash box (y).
2. Altruism, self-image and/or warm glow of giving, i.e. utility derived from transferring money to the charity.[14]
3. The social image of A with respect to the observing player(s) j ($SI_{A,j}$), which only comes into play if x is observable by j.

Consequently, we get:

$$U_A = U(100 - x,\ x,\ E(y|x),\ SI_{A,j}(x)) \tag{1}$$

For expositional purposes, we assume that the utility function is additively separable in its arguments (all results also hold for the general case):

$$U_A = u(100 - x) + v(x) + w(E(y|x)) + SI_{A,j}(x) \tag{2}$$

where u, v, and w are strictly increasing and u is strictly concave. A central question in our design is the relation between SI and x. When modeling SI, we follow Bursztyn and Jensen (2017) who refine the framework of Bénabou and Tirole (2006). Suppose there are two types of subjects, which we label altruistic (a) and selfish (s). Then SI is given by

$$SI_{A,j}(x) = \lambda_{A,j} E_A(\omega_j(a)) Prob_j(\sigma_A = a|x) \tag{3}$$

In this formulation $Prob_j(\sigma_A = a|x)$ denotes the probability that j thinks A is of type

[14] Since the decisions only from Player A will decide the payoff to the charity (due to "pay one" nature of incentives implemented for the reasons discussed above), we cannot distinguish between altruism and self-image or a warm glow of giving. Such a distinction is, however, not the focus of our paper.

a given the observed value of x. $E_A(\omega_j(a))$ gives A's expectations of the social desirability of being seen as type a by group j and, finally, $\lambda_{A,j}$ measures to which extent A cares about her image in group j.

Let us first consider the *SelfSig* treatment. In this treatment, no other player can observe x. Thus, neither y nor SI are influenced by Player A's actions. The first-order condition (FOC) for optimal x becomes

$$u'(100 - x) = v'(x) \tag{4}$$

i.e., in the optimum, the marginal loss in utility due to decreasing the private account equals the marginal utility from altruism. Comparatively, in the *SocSig* treatment x is observed by Player C, meaning that Player A's actions are influenced by her self signaling and social image but not by reputation concerns. The corresponding FOC becomes

$$u'(100 - x) = v'(x) + SI'_{A,j}(x) = v'(x) + \lambda_{A,j} E_A(\omega_j(a)) \frac{d}{dx} Prob_j(\sigma_A = a|x) \tag{5}$$

It seems reasonable to assume that $Prob_j(\sigma_A = a|x)$ is increasing in x. If we assume that both $\lambda_{A,j}$ and $E_A(\omega_j(a))$ are positive, we can hypothesize that the optimal x in *SocSig* is higher than in *SelfSig*, since u is concave.

Hypothesis 1: *The average amount transferred to the charity should be higher in SocSig than in SelfSig.*

This stated, notice that, if either $\lambda_{A,j}$ or $E_A(\omega_j(a))$ are negative, we could observe a lower x in *SocSig*. In the *Reputation* treatment, only player B can observe x. Here, besides SI, y can also be influenced by x. We get the FOC

$$\begin{aligned} u'(100 - x) &= v'(x) + w'(E(y|x)) \frac{d}{dx} E(y|x) + SI'_{A,j} \\ &= \lambda_{A,j} E_A(\omega_j(a)) \frac{d}{dx} Prob_j(\sigma_A = a|x) \end{aligned} \tag{6}$$

Since $E(y|x)$ should depend positively on x, we get the following hypothesis:

Hypothesis 2: *The average amount transferred to the charity should be higher in Reputation than in SocSig.*[15]

[15] An alternative assumption would follow the crowding-out literature (Bénabou and Tirole, 2006, for experimental findings also see Frey and Oberholzer-Gee, 1997; Gneezy and Rustichini, 2000a). This would suggest that individuals donate because of altruistic reasons (and potential social image concerns) in SocSig,

Behavior in SelfSig measures pure altruism. Comparing the optimal x in Treatment 2 with that in the SelfSig informs us about the role of social image in promoting pro-social (giving) behavior. In comparison to *SelfSig*, *Reputation* captures both social image and pure reputation concerns, and therefore resembles indirect reciprocity. Finally, the impact of pure reputation concerns is captured by the difference of x in *SocSig* and *Reputation*. Note that as predicted by the model of Bénabou and Tirole (2006), monetary incentives as in our *Reputation* treatment can reduce intrinsic motivation and, therefore even lead to a net crowding out of pro-social behavior (Frey and Oberholzer-Gee, 1997; Gneezy and Rustichini, 2000a). Theoretically (Bénabou and Tirole, 2006) and empirically (Gneezy and Rustichini, 2000b), such a net crowding-out only occurs if incentives are sufficiently low or wiggle-room sufficiently high (Ariely et al., 2009; Linardi and McConnell, 2011; Exley, 2015). The multiplier in our setup ensures that incentives to do good are high and amplified further when behavior contains signaling value, as is the case in our *Reputation* treatment. Thus, in line with Hypothesis 2, a net-crowding out seems unlikely.

In our analysis, instead of the precise magnitude of x, its sign is of interest. While negative values of x could certainly be regarded as a social norm violation, $x = 0$ is the default, and any positive values of x should boost both SI and y. Therefore, we also formulate a hypothesis with respect to observed frequencies of certain types of behavior.

Hypothesis 1*: *The frequency of positive (negative) values of x should be higher (lower) in SocSig than in SelfSig.*

Hypothesis 2*: *The frequency of positive (negative) values of x should be higher (lower) in Reputation than in SocSig.*

3.4. Results from the Laboratory Experiment

Our design allows us to examine the effects of reputation concerns, self-signaling, and social signaling both on the intensive and extensive margin. We first investigate the intensive margin, examining the *magnitude* of shifts in charitable giving, and then the extensive margin, in this case the *frequency* of pro-social (giving) and anti-social (taking) behavior. Unless otherwise specified, mean test p-values result from two-sided non-parametric tests.

whereas behavior in the Reputation condition is driven by monetary concerns, too. We instead follow the model predictions as outlined above. Our results support this empirically, as will be illustrated shortly.

Magnitude of Behavior

In this section, we attempt to answer: do the sums of money that Player A took away from or gave to charity differ across treatments? Figure 4 displays the net percentage change in the charity account due to Player A's actions. A positive (negative) number indicates a net benefit (net loss) for the charity.

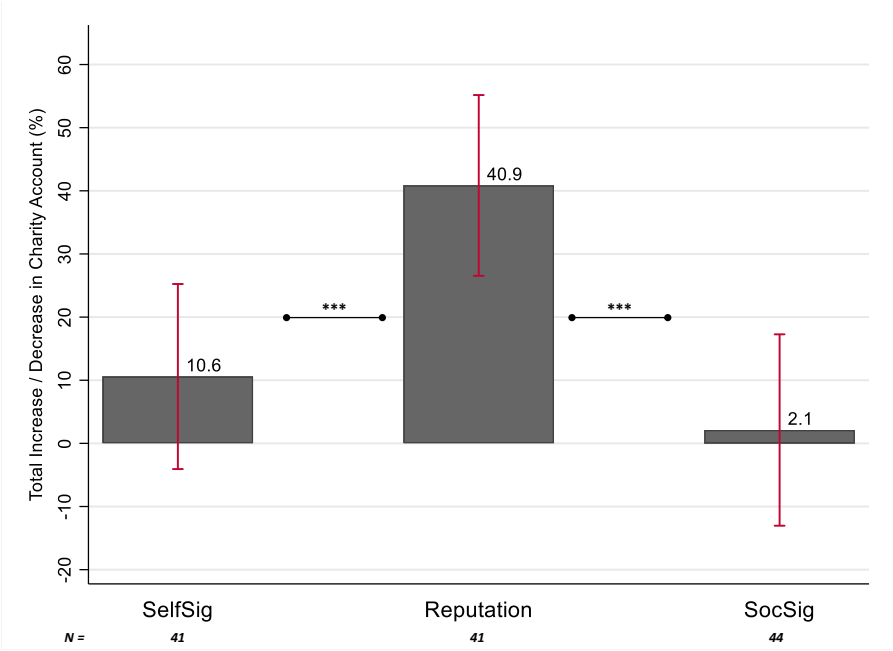

Figure 4: Magnitude of change in charity account (compared to initial endowment) across treatments. Observations per treatment are displayed at the bottom of each column. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels. Whiskers represent 95% CIs.

Consistent with what was anticipated in our hypotheses section, the largest spike of pro-social behavior appears in the treatment in which Player A's behavior was directly observed by Player B (*Reputation*). Here, the charity gained on average significantly more than in the self-signaling (SelfSig) or social signaling (SocSig) treatments. Mann-Whitney-U (MWU) test statistics indicate a highly significant difference between *Reputation* and the two other treatments (p<0.01 and p<0.001, respectively). Surprisingly, however, is that net contributions fall with the introduction of social signaling; 10.6% SelfSig versus 2.1% SocSig. While this difference is not statistically significant (p=0.61), the direction of

the change is in contrast to the theoretical predictions. Figure 5 displays coefficient plots in which we examine total behavior, taking behavior, and giving behavior separately.[16]

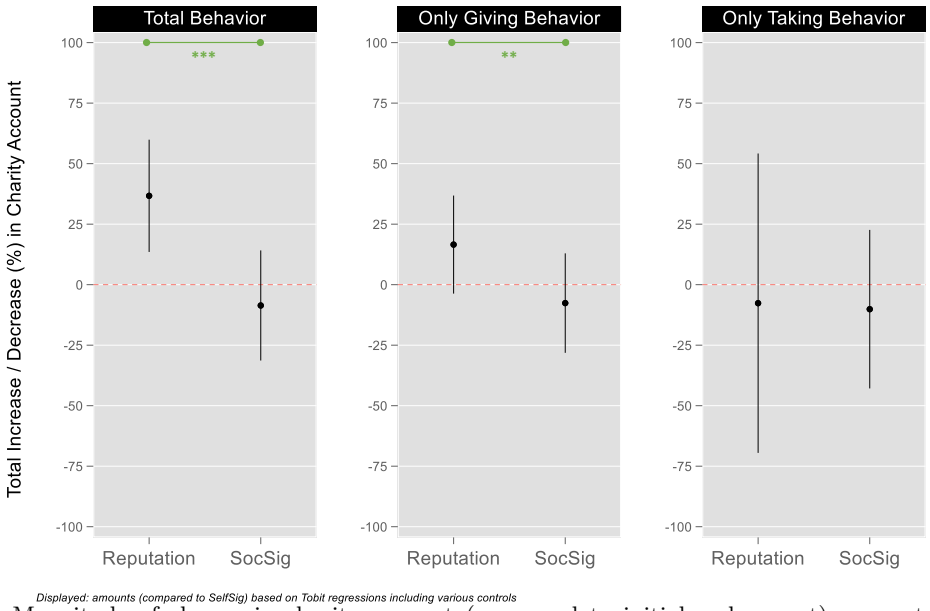

Figure 5: Magnitude of change in charity account (compared to initial endowment) across treatments. Coefficients (compared to SelfSig) based on Tobit regressions including various controls. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels. Whiskers represent 95% CIs.

Across multiple specifications, we find significantly different behavior in the Reputation treatment, which is consistent with our non-parametric results.[17] Note that in line with our previous finding, the post-estimation test between Reputation and SocSig treatment suggests a significantly lower amount of money given to charity, both for overall behavior as well as giving behavior only. We do not observe a significant difference for taking behavior.

Frequency of Behavior

Next, we examine individual heterogeneity with respect to the extensive margin: distribution of behavior categories (taking, no change, and giving) and break it out across

[16]The coefficient plots are a graphical illustration of the Tobit regression analysis including several covariates, which we present in the Appendix A.1 and illustrate them in Figure A.2.

[17]For total change and giving behavior, our results in Appendix Table A.1 also suggest a significantly positive link between Player A's behavior and her beliefs about the extent of reimbursement Player B will provide through the transfer from the cash box, leading to significantly more giving when higher reimbursement is anticipated.

treatments in Figure 6. Compared to the SelfSig, pro-social behavior is significantly more frequent and anti-social behavior is significantly less frequent in the *Reputation* treatment, which is in line with our Hypotheses 1* and 2*. Consistent with our previous results on the magnitude of behavior, the directional change of taking behavior in SocSig compared to SelfSig – though insignificant – is surprising and runs counter to our hypotheses.

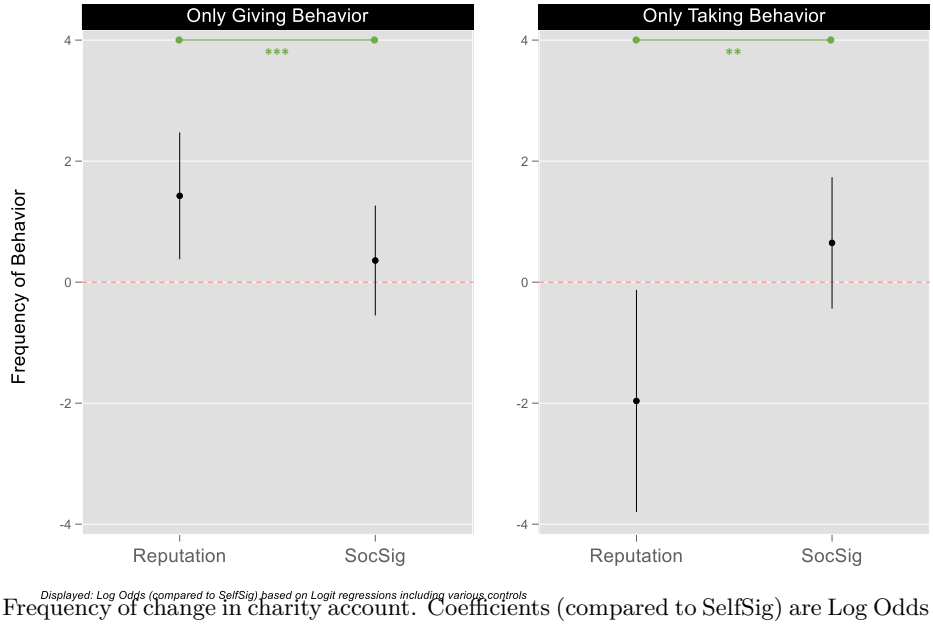

Figure 6: Frequency of change in charity account. Coefficients (compared to SelfSig) are Log Odds and based on Logit regressions including various controls. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels. Whiskers represent 95% CIs.

In sum, the existence of reputation concerns proved an effective means to increase (decrease) pro-social (anti-social) behavior with respect to the frequency at which such behavior occurs. This is in line both with the previously discussed literature and our hypotheses section. At the same time, contrary to initial expectation, the presence of social signaling failed to increase contributions, and descriptively, although not statistically, seemed to facilitate anti-social behavior compared to the SelfSig condition. Our working assumption is that our design can produce two countervailing effects simultaneously that work in opposite directions: existing social image concerns (leading to welfare-enhancing behavior) versus the inability to persuade Player B to transfer money from the cash box in return for Player A's pro-social behavior towards the charity (leading to welfare-damaging behavior). These findings motivate our subsequent experiments as discussed in Section 4.

4. Follow-Up Experiments on MTurk

Our analysis so far suggests that relying exclusively on social incentives alone through observability, without monetary repercussions, produces little in the way of aggregate social benefit. On the other hand, we find that, paired with economic incentives, observability of one's actions triggers the expected reputation concerns and produces welfare-enhancing behavior. The second finding, however, suggests a potential confound to the first result: as previously discussed, much of Player A's giving behavior is driven by her beliefs regarding Player B's reciprocity (reimbursement through the cash box). In other words, the behavioral change in the *SocSig* treatment can be interpreted as the result of two countervailing forces pulling simultaneously into opposite directions: existing social image concerns (leading to welfare-enhancing behavior) versus the inability to persuade Player B to transfer money from the cash box in return for Player A's pro-social behavior towards the charity (leading to welfare-damaging behavior).

To study the pure effect of the observability nudge in the presence of social image concerns, we simplify the design by closing down the latter channel. We do so by removing the cash box and Player B altogether and focus on the comparison between *SelfSig* and *SocSig*. As will be discussed in the results section below in more detail, limiting the channel to observability alone – without the opportunity to be reimbursed through a cash box – leads to a traceable backfiring of this nudge, which is in line with the intuition mentioned above. Results from our refined design indicate a significant *decrease* of the charity endowment compared to a situation without observability (SelfSig), which is in line with the directional results in our laboratory experiment. To get to the bottom of the behavioral mechanism, we administer various additional experiments that allow us to tease out the causes for the backfiring and – importantly from a policy point of view – interventions that mute the backfiring. To achieve this and meet the demands of high-powered studies, we turn to Amazon Mechanical Turk (MTurk) and collect data from over 1,600 participants across 8 treatments.[18] We structure our investigation into three steps:

1. Examining the role of observability where the existence of another participant

[18] Beyond achieving the required statistical power, turning to MTurk has the additional advantage of testing these interventions in a more diverse sample, which has shown to produce robust results compared to laboratory settings (Arechar et al., 2018; Coppock et al., 2018; Snowberg and Yariv, 2018). To ensure high quality data on MTurk, we utilize a combination of CAPTCHAs and screening questions to avoid pool contamination. We applied the following restrictions to the participant pool: participants had to be in the U.S., approval rate was greater than 99% on MTurk, and they had not taken this study before.

(Player B), additional monetary incentives (cash box), and beliefs about such are absent by design. We compare behavior between *SelfSig* and *SocSig* conditions and establish that a backfiring of the nudge exists, which informs step 2.

2. Examining the reasons why the backfiring occurs. We borrow from the social preference literature on inequality aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) and examine how observability shapes inequality concerns. We introduce different versions of both SelfSig and SocSig treatments in which the ex-ante distribution of money between the players is varied while keeping the endowment of the charity and the overall wealth (sum of all endowments) constant. Our results confirm that this affects the extent of backfiring in the theoretically predicted direction, which informs step 3.

3. Examining interventions that can mute the backfiring. We borrow from the social norms literature (Cialdini et al., 1990, 1991, as implemented by Krupka and Weber, 2009) and introduce variations of both SelfSig and SocSig treatments that use simple norm-focus interventions. Our results confirm that these interventions work in the expected direction.

4.1. Design

In order to be able to examine the pure effect of observability in *SocSig* beyond what was possible up to now, we removed Player B altogether and focused solely on Player A's behavior towards a charity with (SocSig) and without (SelfSig) observation by a fully passive Player C. In both treatment variations, participants were truthfully informed that there is also another MTurk participant with the respective endowment who, depending on the treatment variation, either would or would not be able to observe Player A's decision towards a charity. Figures A.4 and A.5 in the Appendix illustrate the basic design features in detail. By removing Player B, we retain a clean design to focus on the behavioral impact of social signaling in these two treatments and allows us to examine whether the previously observed ineffectiveness is due to no effect or countervailing forces.[19] Results in Section 4.4.1 suggest that observability without monetary repercussions can indeed backfire.

[19] One implication from this is that, in expected value, A's income is reduced because no reimbursement through B's Cash Box will be possible anymore. This design choice has the advantage that we limit the impact that the observability through C can have on A's behavior to the necessary minimum, allowing us to focus on a clean identification of the observability effect.

With these results in mind, we design additional treatments that investigate the source of this backfiring. In line with Fisman et al. (2017) and Exley and Kessler (2018) and using insights from Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), we test the reasonable assumption that the focus on this inequality increases if player C can observe the action of player A, thereby creating the backfiring effect. To test this, we implement two additional variations of the SelfSig and SocSig in which the ex-ante payoffs of Players A and C are systematically varied (while keeping the overall wealth – sum of all endowments – constant to achieve maximum comparability) to achieve either advantageous or disadvantageous inequality (from the perspective of Player A). In particular, our two variations include *advantageous* inequality (AI), in which player A starts with a substantially higher endowment than player C, and *disadvantageous* inequality (DI), in which it is the opposite.[20] Our results in Section 4.4.2 suggest that observability indeed affects the salience of inequality perceptions and can explain the backfiring.

Finally, we design additional treatments that investigate how to limit the backfiring of the observability nudge. In concordance with Focus Theory (Cialdini et al., 1990, 1991), we expect that manipulating the strength of the norm focus, by drawing attention to the norm, can render the social signal nudge more effectively (for a discussion of when and how norm focus is effective, see Kallgren et al. 2000; Bicchieri and Dimant 2019). Our treatment variation follows the focus intervention designed by Krupka and Weber (2009): we elicited incentivized beliefs (the outcome of which was only revealed at the very end of the experiment) about the behavior of other participants right before participants were able to make their decision with respect to giving to or taking from the charity. The belief elicitation aimed at putting more focus on the norm by asking participants what they thought other participants in the same situation had done previously (empirical expectations). As Krupka and Weber (2009) have shown, such a belief elicitation can increase giving in a dictator game and is thus the right approach for our setup (a variant of a dictator game). We refer to these treatments as *SelfSig BE* and *SocSig BE*, respectively.

[20] We refer to these treatments as *SelfSig AI*, *SelfSig DI*, *SocSig AI*, and *SocSig DI*, respectively. For both DI and AI, we choose the initial endowments such that player A can surpass player C if she decides to make sufficient adjustments to the endowment of the charity via giving or taking. Importantly, however, we keep both the endowment of the charity as well as the total welfare (sum of all endowments) the same compared to the previous treatments and constant across all new conditions to allow for maximum comparability.

4.2. Participants and Procedures

The one-shot between subjects design of our treatments mirrors the design implemented in our lab experiment. As indicated in Table 2 below, we collected data for a total of 1,650 participants in the role of Player A (49.4% female), and the average age was 33.7 years.[21]

Treatments	Endowment (A / Charity / C)	C Observes A?	# Player A
SelfSig	100 / 100 / 100	No	226
SelfSig BE	100 / 100 / 100	No	188
SelfSig AI	180 / 100 / 20	No	217
SelfSig DI	80 / 100 / 120	No	183
SocSig	100 / 100 / 100	Yes	221
SocSig BE	100 / 100 / 100	Yes	184
SocSig AI	180 / 100 / 20	Yes	223
SocSig DI	80 / 100 / 120	Yes	208

Table 2: Descriptive Statistics for Player A. *BE* refers to *Belief Elicitation* treatment, *AI* to *Advantageous Inequality* treatment, and *DI* to *Disadvantageous Inequality* treatment.

Participants and the respective charities started with an endowment of \$1 each and the payment mechanism remained the same as in Experiment 1 with the exception that Player B was not present anymore. Correct beliefs were incentivized with \$0.5. From start to finish, the duration of a treatment was 10 minutes. This yielded an average hourly income of \$7.72 for Player A, well above MTurk standards (Hara et al., 2018).[22]

4.3. Extended Theory and Behavioral Predictions

In this section we will extend our previous theory in order to account for a potential reason of backfiring, i.e. the interaction of observability with inequality aversion. Note that the decision of player A to give money to the charity or take money from it also introduces inequality with respect to player C. Initially there is equality between A and C as both have an endowment of 100. In our new design, taking (resp. giving) of A now introduces (dis)advantageous inequality with respect to C. In contrast, in our laboratory experiment

[21] As before, all our analyses focus on the behavior of Player A.

[22] We find that MTurk participants are significantly older than our lab sample ($p<0.01$) and a more equal male to female ratio ($p=0.02$). We do not find any significant differences in risk-taking behavior ($p=0.14$). Overall, MTurk participants are less charitable than our lab participants. For a comprehensive discussion of lab versus MTurk populations, see Snowberg and Yariv (2018).

the effect of giving/taking on inequality was ambiguous because there was additionally the unknown amount received from the cash box. Since the work of Fehr and Schmidt (1999) and Bolton and Ockenfels Bolton and Ockenfels (2000) it is obvious that inequality aversion is a strong driver of behavior. Therefore, it seems reasonable that also in the present experiment inequality can influence giving/taking behavior of A. Since A ends up with $100 - x$ and C with 100, inequality is only determined by x.

As we vary initial endowments of A and C in some treatments let us denoted them by E_A and E_C respectively. Concentrating on the pure effect of inequality and following the model of Fehr and Schmidt (1999) the utility of A is then given by

$$U_A = E_A - x - \alpha \max(E_C - E_A + x, 0) - \beta \max(E_A - x - E_C, 0), \tag{7}$$

where the parameter α (β) measures the distaste of disadvantageous (advantageous) inequality of A.

In our experimental design, this pure effect of inequality aversion is complemented by the additional motivations identified in the previous model, i.e. self-image (resp. altruism or warm glow of giving) and social image.[23] Restricting attention to the additive-separable variant of the utility function we thus get

$$U_A = u(E_A - x) + v(x) + SI_{A,j}(x) - \alpha \max(E_C - E_A + x, 0) - \beta \max(E_A - x - E_C, 0). \tag{8}$$

This leads to the following first-order conditions:

$$u'(E_A - x) = v'(x) + SI'_{A,j}(x) - \alpha \quad \text{if} \quad E_C > E_A - x \tag{9}$$

and

$$u'(E_A - x) = v'(x) + SI'_{A,j}(x) + \beta \quad \text{if} \quad E_C < E_A - x \tag{10}$$

Let us assume as before that $SI'_{A,j}(x) > 0$ and consider the case of our main treatments where $E_A = E_C$. Then the first (second) FOC is relevant for positive (negative) values of x. This shows that for subjects who give (take) a higher degree of inequality aversion induces them to give (take) less.

[23] Since player B is absent in the new design, the effect of x on the amount received from the cash box becomes irrelevant.

As explained in Section 4.1, work of Fisman et al. (2017) and Exley and Kessler (2018) suggests that observability may enhance inequality concerns. We hypothesize that this is particular true for disadvantageous inequality aversion: having less than another person is more painful if the other person also knows this. For advantageous inequality this effect is less clear as some people may actually like to show off that they have more than the other person. Therefore, we assume that α is higher in SocSig than in SelfSig where as β remains unchanged. This leads to the following hypotheses.

Hypothesis 3: *The average amount transferred to the charity is higher in SelfSig than in SocSig if inequality aversion dominates social image concerns.*

Hypothesis 4: *The average amount transferred to the charity in SocSig is higher if the initial endowment of A is increased relative to that of C.*

The last hypothesis follows simply from the fact that if A has a higher endowment than C she can give something to the charity and still be better than C, i.e. the effect of disadvantageous inequality does not come into play. Therefore, the backfiring effect has no or less effect in this case. Again, we can formulate both hypothesis also with respect to frequencies.

Hypothesis 3*: *The frequency of positive (negative) values of x is higher (lower) in SelfSig than in SocSig if inequality aversion dominates social image concerns.*

Hypothesis 4*: *The frequency of positive (negative) values of x in SocSig is higher (lower) if the initial endowment of A is increased relative to that of C.*

4.4. Results from the MTurk Experiments

4.4.1. Refined Examination of Observability Without Monetary Repercussions

First, we perform a comparison between SelfSig and SocSig where both Player B and the additional cash box are absent. In addition to finding that giving behavior reduces significantly in SocSig compared to Selfsig, the key result is that SocSig leads to a significant backfiring effect overall with respect to money remaining in the charity account (-28.3% vs. -18.4%, MWU, p=0.0401). This supports both Hypotheses 3 and 3*. In light of our original findings where additional monetary rewards (and the corresponding beliefs about them) were present in the form of Player B and the cash box, this result suggests that Player A's

behavior is mainly driven by self-serving motives and is in fact amplified by observability when additional countervailing forces in form of monetary incentives are absent.[24]

Result: *The social signaling nudge alone (SocSig) backfires compared to behavior observed in SelfSig.*

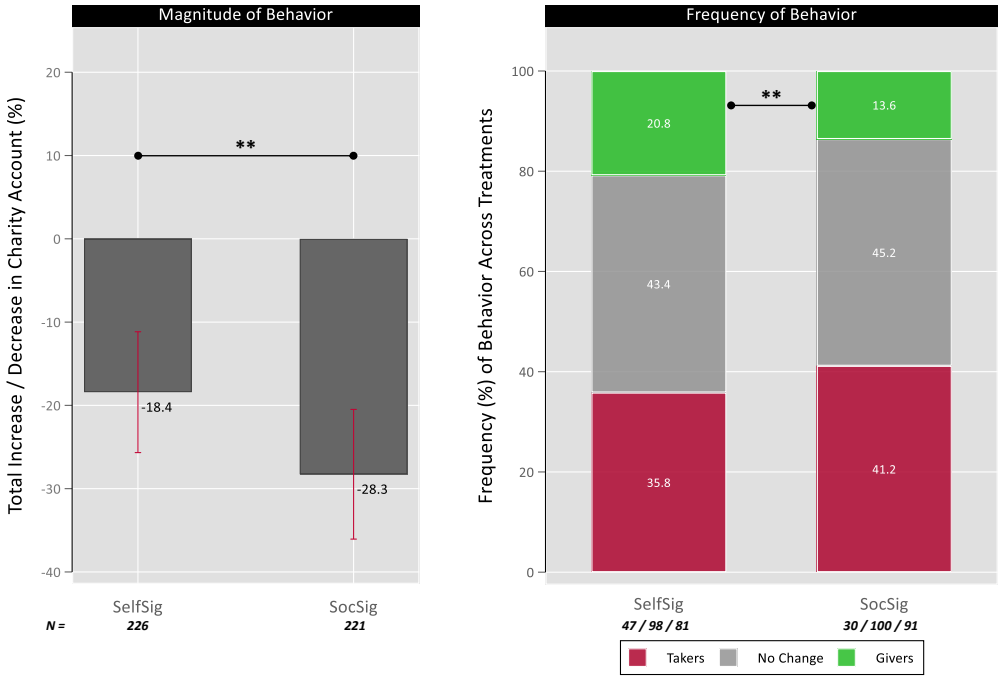

Figure 7: Magnitude (left panel) and frequency (right panel) of change in charity account (compared to initial endowment) across treatments. Observations per treatment are displayed at the bottom. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels. Whiskers represent 95% CIs.

We will unpack this finding in the next chapters, explain the underlying mechanism for this effect, and provide a solution to mute it.

4.4.2. Explaining the Backfiring Effect: The Role of Inequality Concerns

In light of the results, we first examine whether the originally observed backfiring effect of SocSig compared to SelfSig can be explained by a cross effect between inequality aversion and observation. Fisman et al. (2017) and Exley and Kessler (2018) show that equity

[24] We present the distribution of behavior across treatments in the Appendix in Figure A.9.

concerns can be narrowly bracketed, more focused on some dimensions of a distribution problem than on others. Hypotheses 3 and 3* extend this bracketing effect to observation and posit that, for some deciders, knowing that the observer can see their decision triggers a social comparison effect. Their deliberations turn away from how much they should give to charity and towards the social comparison of their own payoff versus the observer's payoff. In effect, they may now perceive the situation as a dictator game (rather than a charitable giving game) in which the receiver's payoff is fixed but the decider can increase or decrease one's own payoff. Distribution-based social preference models would then lead us to expect that deciders whose attention is reoriented in this way would then act to achieve relative distributions between self and observer of the kind we observe in the dictator game. This would explain the reduction in giving, including the increase in taking, observed in SocSig relative to SelfSig. According to Hypotheses 3 and 3* giving goes down in SocSig relative to SelfSig because observation leads Player A to become more focused on payoff inequality between herself and Player C. The new experiment provides a validation test of the hypotheses in previously unexplored circumstances. We collect data along two variations of both SelfSig and SocSig treatments of the original experiment:

1. **Advantageous Inequality (AI):** Player A starts with a larger endowment (180) than Player C (20).

2. **Disadvantageous Inequality (DI):** Player A starts with a smaller endowment (80) than Player C (120).

In AI, player A reduces inequality with respect to C by giving to the charity. If the focus on inequality is stronger in SocSig, we expect a higher total amount in the charity box than in SelfSig. In contrast, in DI inequality is reduced by taking from the charity. Here, we expect in SelfSig a higher total amount in the charity box than in SocSig (see Hypotheses 4 and 4*). Analogous to the previous analyses, we present our results both in terms of magnitudes (Figure 8 below) and frequencies (Figure A.6 in the Appendix).

Our results support the consideration that backfiring is caused by inequality aversion: under AI, we observe a statistically significantly larger amount left in the charity account in SocSig compared to SelfSig (-8.3% vs. -21.9%, MWU, p=0.0308). Consistent with this, the charity account is substantially more depleted under DI in SocSig compared to SelfSig (-34.9% vs. -19.8%, MWU, p=0.0154). The depletion in SocSig DI is also substantially

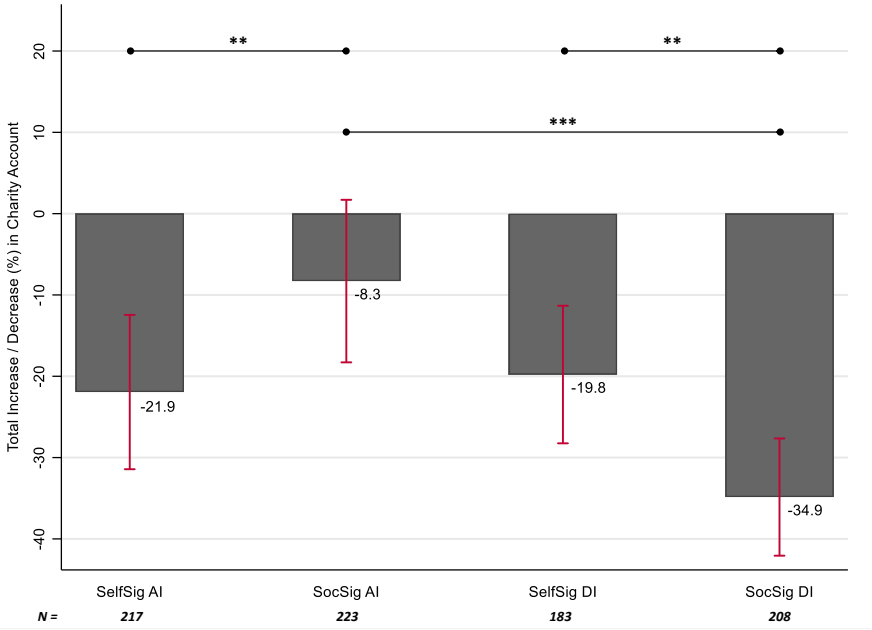

Figure 8: Magnitude of change in charity account (compared to initial endowment) across treatments. Observations per treatment are displayed at the bottom of each column. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels. Whiskers represent 95% CIs.

and highly significantly larger than in SocSig AI (-34.9% vs. -8.3%, MWU, p<0.001).[25]

Overall, our results strongly suggest that the original backfiring effect can be explained by inequality aversion based on the idea that the observation alters the focus of the (in)equality reference. The regression coefficients in Figure 9 corroborate these findings. Our results on the *frequency* of behavior change (Figure 10) are consistent with the previous analyses and our story: compared to SelfSig, in SocSig advantageous inequality leads to significantly more giving and significantly less taking behavior, while disadvantageous inequality reverses the effect completely. We can conclude that the observed backfiring effect found in SocSig is indeed mediated by the observed inequality between both players.[26]

[25] These changes in the charity box are complemented by respective changes in the composition of takers and givers, as indicated in Figure A.6 in the Appendix.

[26] Presumably, even when payoffs are equal a priori (as was the case in the standard variant of the experiment), Player A's drive to signal payoff superiority to Player B is still present. One possible reason is that Player A is under the impression to deserve more because she is the one doing the work.

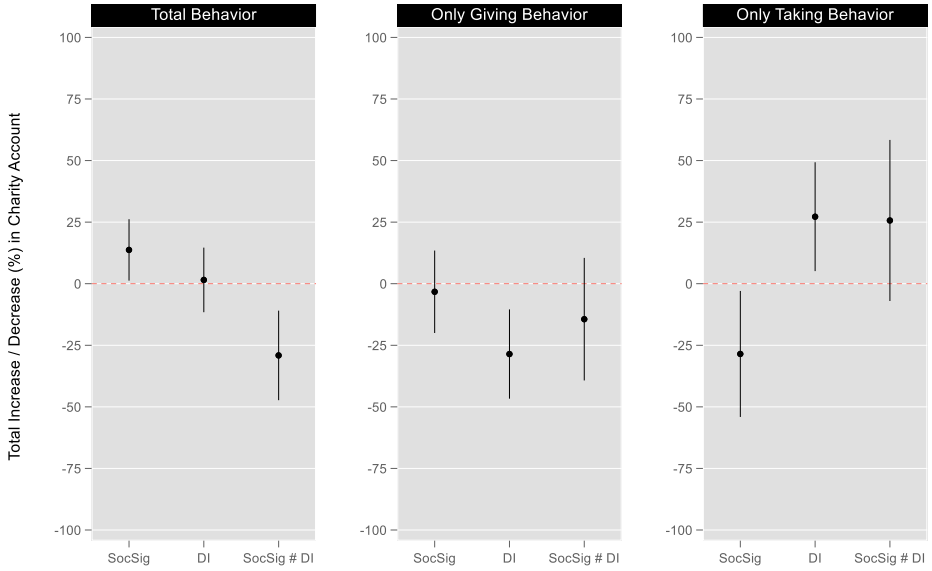

Figure 9: Magnitude of change in charity account. Coefficients (compared to SelfSig) based on Tobit regressions including various controls. Whiskers represent 95% CIs. Full regression full output presented in Table A.3 in the Appendix.

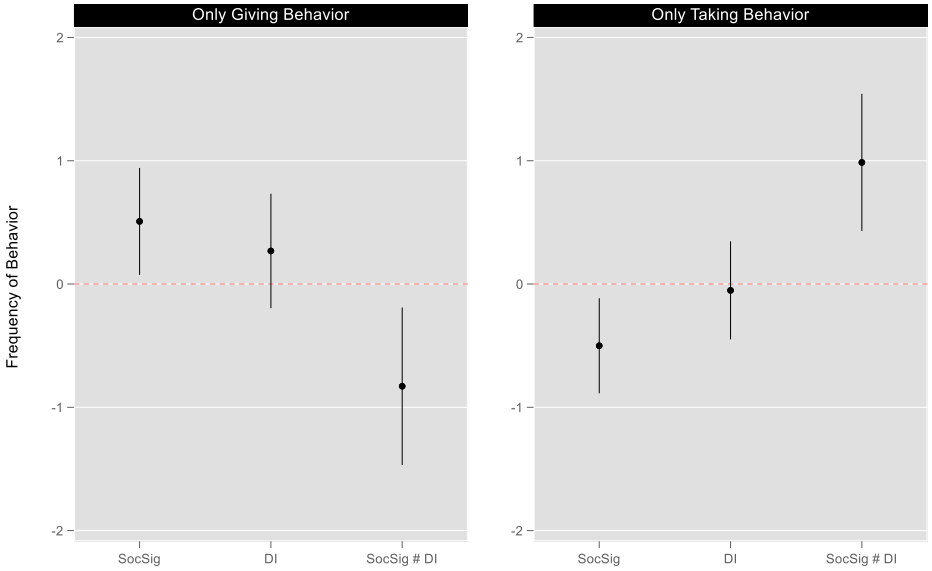

Figure 10: Frequency of change in charity account. Coefficients (compared to SelfSig) are Log Odds and based on Logit regressions including various controls. Whiskers represent 95% CIs. Full regression full output presented in Table A.4 in the Appendix.

4.4.3. Resolve the Backfiring: The Role of Norm Focus

Finally, we provide insights into one behavioral intervention capable of resolving the observed backfiring effect that is guided by the norm focus literature (Cialdini et al., 1990, 1991; Krupka and Weber, 2009; Bicchieri and Dimant, 2019), as discussed in Section 4.1. In line with our previous analyses, we split the analysis into two parts: Magnitude of contribution and frequency of pro- and anti-social behavior for all four treatments (SelfSig and SocSig with and without belief elicitation). We present the beliefs for the two belief elicitation (BE) treatments in the Appendix in Figure A.7.

Magnitude of Behavior

First, we examine the intensive margin of behavior as presented in Figure 11.

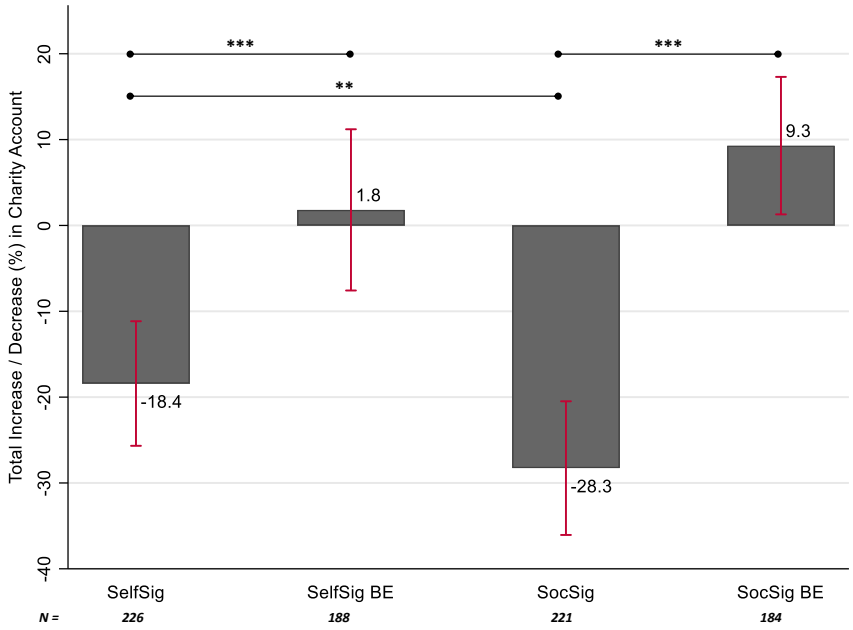

Figure 11: Magnitude of change in charity account (compared to initial endowment) across treatments. Observations per treatment are displayed at the bottom of each column. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels. Whiskers represent 95% CIs.

Within each condition, we observe that the norm focus intervention substantially increases (decreases) the monetary amounts given to (taken from) charity, both in the SelfSig (1.8% vs. -18.4%, MWU, p<0.001) as well as the SocSig (9.3% vs. -28.3%, MWU, p<0.001)

treatments. These findings suggest that a norm focus through the use of belief elicitation can revert the anti-social tendencies of individuals. We do not observe the norm focus to exert a differential impact on SocSig (9.3% vs. 1.8%, MWU, p=0.392).

Result: *The norm focus intervention works in the expected direction in the form of increasing (decreasing) the magnitude of giving (taking) behavior.*

The findings suggest that a reminder to think about what other people do creates the expected behavioral reaction, suggesting that the focus shifts away from purely self-serving gain towards social welfare considerations (for related findings, see Schultz et al., 2007).[27] We present the coefficients of the regression output graphically in Figure 12 and corroborate our results through a series of Tobit regressions both for total change and for taking and giving behavior separately.

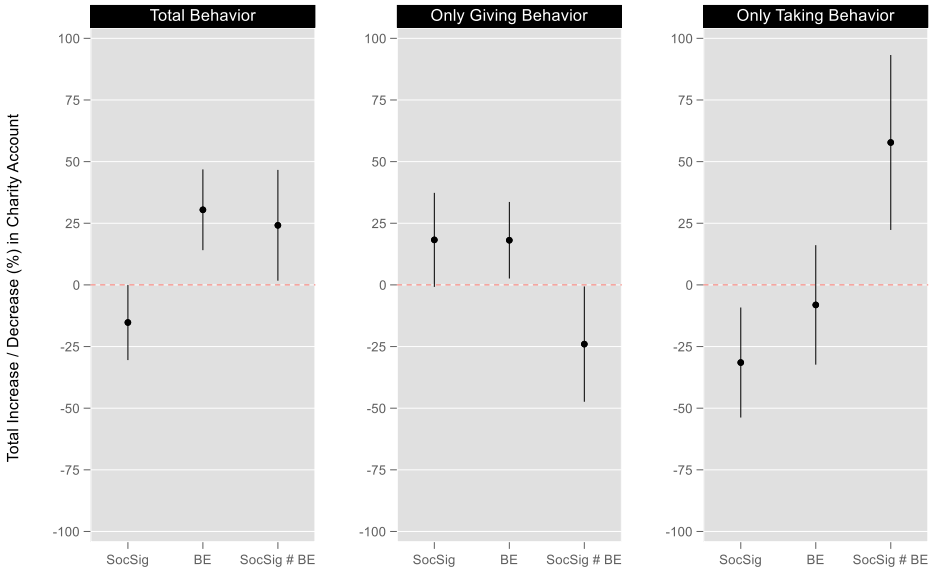

Figure 12: Magnitude of change in charity account. Coefficients (compared to SelfSig) based on Tobit regressions including various controls. Whiskers represent 95% CIs. Full regression full output presented in Table A.5 in the Appendix and a graphical breakdown by amount given and taken in Figure A.8.

[27] There, the authors describe a boomerang effect of norm interventions that capitalize on a descriptive normative message involving telling individuals the exact extent of an undesired behavior (excessive energy consumption) of their peers. Exposing the actual behavior produced either desirable energy savings or an undesirable boomerang effect, depending on whether the individuals in this condition were initially consuming at a low or high rate.

Strikingly, these results suggest that such an intervention can backfire even without exposure to actual peer behavior, which can at least partially be attributed to the beliefs the participants hold about the behavior of others (see previous discussion of our laboratory experiment and Figure A.7 in the Appendix). The mere focus on what behavior *could be* is seemingly sufficient to backfire (for related findings in the context of trust and lying behavior, see Bicchieri et al., 2019c,b; Dimant et al., 2019).

Frequency of Behavior

The first noteworthy finding is that the initial surprising result from Experiment 1 reproduces, and now yields, a statistically significant difference between SelfSig and SocSig in the form of *reduced* giving behavior frequency in SocSig.[28] We do not observe any significant differences for taking behavior. We also find that the norm focus intervention is particularly effective in increasing (decreasing) giving (taking) behavior for both SelfSig and SocSig compared to when the norm focus is absent. In fact, the interaction term in Figure 13 indicates that the norm focus intervention is over-proportionally effective in the SocSig treatment. We can derive the following two results:

Result 1: *In line with results in Experiment 1, the social signaling nudge alone (SocSig) is insufficient in facilitating giving behavior or reducing taking behavior. In fact, social signaling even backfires and substantially reduces giving behavior.*

Result 2: *The norm focus intervention works in the expected direction in the form of increasing (decreasing) the frequency of giving (taking) behavior.*

Importantly, one potential explanation for these results concerns the existence of anti-social norms in the MTurk population. If true, not only could the default behavior expected to be *taking from the charity*, but the observation by others (as is the case in SocSig) should be expected to amplify this effect. We use the incentive-compatible Krupka and Weber (2013) method to establish that this is not the case. [29]

Taken together, these results allow us to harmonize the main finding of backfiring.

[28] This finding can be harmonized not only with higher statistical power, but also with respect to the Focus Theory: the absence of Player B leads to a reduced focus, in particular in the SocSig condition, allowing for the difference with SelfSig to be more strongly pronounced than in Experiment 1.

[29] For this, we ran a questionnaire with a separate pool of 200 participants on MTurk in which participants were explained the context and action space of the experiment and had to guess the modal response to each behavior (taking from charity, not changing equal split, giving to charity) on a 4-item scale (very socially inappropriate, socially inappropriate, socially appropriate, very socially appropriate). As Figure A.11 in the Appendix indicates, giving is clearly the norm, whereas taking form the charity is despised.

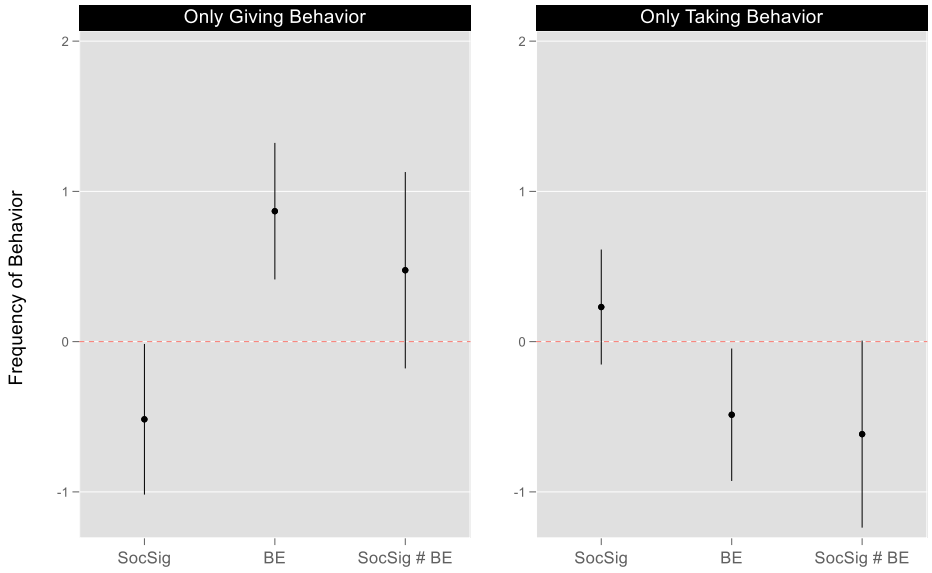

Figure 13: Frequency of change in charity account. Coefficients (compared to SelfSig) are Log Odds and based on Logit regressions including various controls. Whiskers represent 95% CIs. Full regression full output presented in Table A.6 in the Appendix.

These are particularly important insights from a policy perspective and we will return to its discussion in our conclusion.

5. Conclusion

Individuals commonly engage in behaviors that constitute a trade-off between self-serving and other-regarding goals. Typically, the trade-offs a person makes are subject to observation by peers. As such, this embodies concerns related to social signaling, self signaling, and future reciprocity. The extent to which there is observation falls under the umbrella term of nudges and is frequently employed to produce behavioral change because observability is argued to affect the degree of plausible deniability both directly (through economic incentives following from reputation concerns) and indirectly (through social incentives following from image concerns) (Thaler and Sunstein, 2008; Von Hippel and Trivers, 2011; Rogers et al., 2018; Bicchieri and Dimant, 2019). Empirically, however, disentangling these channels is difficult because they rarely appear in isolation. Overwhelmingly, past studies of observability have focused on its impact on decisions in the pro-social domain alone (Bradley et al., 2018). To make matters more complicated, the

trade-offs involved are often embedded in strategic environments in which individual and social welfare can be influenced by observers.

Our series of studies aim to address these aspects and advance our scientific understanding of how, why, and when a seemingly simply behavioral intervention to achieve behavioral change does (not) work and – important from a policy point of view – what to do about it. For this reason, we study the impact of an observability nudge in both the pro- and anti-social domain and systematically dissect the different channels at play. We capitalize on both a laboratory and online setting to explore the mechanisms at play in great detail. We complement our experimental analysis with theoretical exercises, which derive two sets of testable hypotheses using models that capitalizes on the following distinct drivers of behavior: self-interest, altruism/warm-glow, social image, and inequality aversion. Broadly speaking, our theoretical analyses predict that reciprocal reputation mechanisms will be most impactful at curbing anti-social behavior and promoting pro-social behavior whereas the benefits of social image observation alone can be eroded and even outmatched by inequality concerns.

The first prediction, that reciprocity will be most successful at reducing anti-social behavior and increasing pro-social behavior is confirmed by our data. This finding is in line with the standard literature on reputation concerns and indirect reciprocity. However, the evidence from our series of experiments lead us to the conclusion that the role of social image concerns in affecting pro- and anti-social behavior is nuanced. Not only does the social image nudge fail to yield the expected benefit, we even find evidence for backfiring.[30]

One possible explanation that can be inferred in recent literature is that one's inequality concerns are a function of one's reference group, which we actively manipulate in our experimental setup (for a discussion see Fisman et al., 2017; Exley and Kessler, 2018). The hypothesis is that by introducing observation some players are reorientated towards social comparison. Players whose attention is not reoriented might then give the same as in SelfSig or actually increase giving as anticipated by the social image hypothesis. But from the data, it appears that the posited reorientation, if it occurs, is strong enough in aggregate to crowd out any aggregate pro-social image effect. Our additional treatments provide independent evidence for this effect of social comparison and inequality concerns.

[30] A number of obvious hypotheses involve factors held fixed in our experiment (economic self-interest, the size of the stakes in the experiment, and experimenter observation effects). While these factors might play a role in some of the observed behavior, they cannot readily explain changes observed across treatments.

In a final step, we turn to another strand of existing literature in an attempt to mitigate the backfiring. Focus Theory (Cialdini et al., 1990, 1991) provides a viable solution to the backfiring effect. Applied to our setup, the theory implies that observation without enforcement can enhance the salience of selfishness, which can be alleviated by increasing the focus on the underlying norm of pro-social behavior. We designed a series of follow-up treatments to examine this hypothesis. In line with the assumptions, observability without monetary reinforcement reduced the amount transferred to the charity significantly. In other words, observation alone produces more extreme behavior in our setup, both on the pro-social and on the anti-social side. But also in line with the Focus Theory hypothesis, when we ask subjects to think about what others would do prior to making their own decisions, we discover that the social observation nudge reverses the backfiring effect and works similarly to observation with monetary concerns. In a way, making a behavioral norm salient resembles the presence of monetary consequences.[31]

Altogether, we can conclude that in situations where both pro- and anti-social behavior is possible, nudges increasing the observability of one's actions have little or even detrimental effects when monetary consequences are absent. Relying on one's social image consideration alone can likely be insufficient, which is consistent with the finding that even extremely 'nice' behavior might not be regarded as such (Klein et al., 2015). Policymakers need to be aware of this potential backfiring effect. Our results can guide them in two respects. First, our results on the role of inequality seem to indicate that the observability nudge should work for relatively well off people, as pro-social behavior here reduces inequality and also may serve as some kind of conspicuous consumption. In contrast, backfiring should be present when the nudge is applied to those who are less well off. Second, we detect a crucial aspect concerning the salience of social norms, which achieves a reversion of the strongly detrimental impact of the pure observability nudge. Our take-away is that in situations in which behavior cannot be enforced or sanctioned, nudges increasing observability of actions should be combined with a salient norm to avoid potential backfiring effects, which – compared to the costly introduction of formal institutions – yields an effective and efficient intervention both in social and economic terms (see results and discussions in, e.g., Bicchieri et al., 2019a; Bicchieri and Dimant, 2019).

[31] Note that unlike findings in, for example, Exley and Kessler (2018), the theory underlying the Norm Focus intervention as it was applied here does not posit a *change* in social norms, but rather *shifting* focus towards already existing norms.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2018). Preferences for truth-telling. *Econometrica*.

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *The Quarterly Journal of Economics*, 94(4):749–775.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Alexander, R. (1987). *The biology of moral systems*. Aldine De Gruyter.

Ali, S. N. and Bénabou, R. (2019). Image versus information: Changing societal norms and optimal privacy. *American Economic Journal: Microeconomics*.

Allcott, H. and Kessler, J. B. (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics*, 11(1):236–76.

Alpízar, F. and Martinsson, P. (2013). Does it matter if you are observed by others? evidence from donations in the field. *The Scandinavian Journal of Economics*, 115(1):74–83.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):464–477.

Andreoni, J. and Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636.

Arechar, A. A., Gächter, S., and Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, 21(1):99–131.

Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? image motivation and monetary incentives in behaving prosocially. *American Economic Review*, 99(1):544–55.

Ayres, I., Raseman, S., and Shih, A. (2013). Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage. *The Journal of Law, Economics, and Organization*, 29(5):992–1022.

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, 126(4):1472–1503.

Balafoutas, L., Nikiforakis, N., and Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, 111(45):15924–15927.

Balliet, D., Wu, J., and De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, 140(6):1556.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11(2):122–133.

Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime*, pages 13–68. Springer.

Behavioral Insights Team (2015). Update report 2013 - 2015. White Paper.

Bénabou, R., Falk, A., and Tirole, J. (2018). Narratives, imperatives and moral reasoning. Technical report, Working Paper.

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

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., Tucker-Ray, W., Congdon, W. J., and Galing, S. (2017). Should governments invest more in nudging? *Psychological Science*, 28(8):1041–1055.

Beshears, J., Choi, J. J., Laibson, D., Madrian, B. C., and Milkman, K. L. (2015). The effect of providing peer information on retirement savings decisions. *The Journal of Finance*, 70(3):1161–1201.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. and Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public Choice*.

Bicchieri, C., Dimant, E., Gaechter, S., and Nosenzo, D. (2019a). Observability, social proximity, and the erosion of norm compliance. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3355028.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2019b). It's not a lie if you believe it: On norms, lying, and self-serving belief distortion. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3326146, University of Pennsylvania.

Bicchieri, C., Dimant, E., and Xiao, E. (2019c). Deviant or wrong? The effects of norm information on the efficacy of punishment. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3294371.

Bohnet, I. and Frey, B. S. (1999). Social distance and other-regarding behavior in dictator games: Comment. *American Economic Review*, 89(1):335–339.

Bolton, G., Greiner, B., and Ockenfels, A. (2013). Engineering trust: reciprocity in the production of reputation information. *Management Science*, 59(2):265–285.

Bolton, G. E., Katok, E., and Ockenfels, A. (2004). How effective are electronic reputation mechanisms? an experimental investigation. *Management Science*, 50(11):1587–1602.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Bradley, A., Lawrence, C., and Ferguson, E. (2018). Does observability affect prosociality? *Proc. R. Soc. B*, 285(1875):20180116.

Brandon, A., Ferraro, P. J., List, J. A., Metcalfe, R. D., Price, M. K., and Rundhammer, F. (2017). Do the effects of social nudges persist? theory and evidence from 38 natural field experiments. Technical report, National Bureau of Economic Research.

Buchan, N. R., Croson, R. T., and Dawes, R. M. (2002). Swift neighbors and persistent strangers: A cross-cultural investigation of trust and reciprocity in social exchange. *American Journal of Sociology*, 108(1):168–206.

Bursztyn, L. and Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, 9:131–153.

Cappelen, A. W., Halvorsen, T., Sørensen, E. Ø., and Tungodden, B. (2017). Face-saving or fair-minded: What motivates moral behavior? *Journal of the European Economic Association*, 15(3):540–557.

Cason, T. N., Friesen, L., and Gangadharan, L. (2016). Regulatory performance of audit tournaments and compliance observability. *European Economic Review*, 85:288–306.

Charness, G., Gneezy, U., and Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, 131:141–150.

Cialdini, R. B., Kallgren, C. A., and Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*, volume 24, pages 201–234. Elsevier.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6):1015.

Coffman, L. C. (2011). Intermediation reduces punishment (and reward). *American Economic Journal: Microeconomics*, 3(4):77–106.

Cooper, D. J. and Kagel, J. H. (2016). Other-regarding preferences. *The handbook of experimental economics*, 2:217.

Coppock, A., Leeper, T. J., and Mullinix, K. J. (2018). The generalizability of heterogeneous treatment effect estimates across samples.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–74.

Damgaard, M. T. and Gravert, C. (2018). The hidden costs of nudging: Experimental evidence from reminders in fundraising. *Journal of Public Economics*, 157:15–26.

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(2):193–201.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Dimant, E. (2019). Contagion of pro- and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*.

Dimant, E., Gerben, A. v. K., and Shalvi, S. (2019). Requiem for a nudge: Framing effects in nudging honesty. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3416399.

Dimant, E. and Tosato, G. (2018). Causes and effects of corruption: what has past decade's empirical research taught us? a survey. *Journal of Economic Surveys*, 32(2):335–356.

Dufwenberg, M. and Muren, A. (2006). Generosity, anonymity, gender. *Journal of Economic Behavior & Organization*, 61(1):42–49.

Ekström, M. (2012). Do watching eyes affect charitable giving? evidence from a field experiment. *Experimental Economics*, 15(3):530–546.

Engelmann, D. and Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*, 67(2):399–407.

Ernest-Jones, M., Nettle, D., and Bateson, M. (2011). Effects of eye images on everyday cooperative behavior: a field experiment. *Evolution and Human Behavior*, 32(3):172–178.

Exley, C. L. (2015). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2):587–628.

Exley, C. L. and Kessler, J. B. (2018). Equity concerns are narrowly framed. Harvard business school working paper 18-040, Harvard University.

Falk, A. and Fischbacher, U. (2002). "crime" in the lab-detecting social interaction. *European Economic Review*, 46(4-5):859–869.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Festré, A. and Garrouste, P. (2015). Theory and evidence in psychology and economics about motivation crowding out: A possible convergence? *Journal of Economic Surveys*, 29(2):339–356.

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

Fisman, R., Kuziemko, I., and Vannutelli, S. (2017). Distributional preferences in larger groups: Keeping up with the joneses and keeping track of the tails.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347–369.

Frey, B. S. and Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *The American Economic Review*, 87(4):746–755.

Gino, F., Ayal, S., and Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological Science*, 20(3):393–398.

Gino, F., Hauser, O. P., and Norton, M. I. (2019). Budging beliefs, nudging behaviour. *Mind & Society*, pages 1–12.

Gneezy, U. and Rustichini, A. (2000a). A fine is a price. *The Journal of Legal Studies*, 29(1):1–17.

Gneezy, U. and Rustichini, A. (2000b). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810.

Grimalda, G., Pondorfer, A., and Tracer, D. P. (2016). Social image concerns promote cooperation more than altruistic punishment. *Nature Communications*, 7:12288.

Hagmann, D., Ho, E. H., and Loewenstein, G. (2019). Nudging out support for a carbon tax. *Nature Climate Change*.

Haley, K. J. and Fessler, D. M. (2005). Nobody's watching?: Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, 26(3):245–256.

Hallsworth, M., Chadborn, T., Sallis, A., Sanders, M., Berry, D., Greaves, F., Clements, L., and Davies, S. C. (2016). Provision of social norm feedback to high prescribers of antibiotics in general practice: a pragmatic national randomised controlled trial. *The Lancet*, 387(10029):1743–1752.

Hamman, J. R., Loewenstein, G., and Weber, R. A. (2010). Self-interest through delegation: An additional rationale for the principal-agent relationship. *American Economic Review*, 100(4):1826–46.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., and Bigham, J. P. (2018). A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 449. ACM.

Hogg, M. and Turner, J. (1987). Social Indentity and Conformity. In *Current Issues in European Social Psychology*, volume 2. Cambridge University Press.

Kajackaite, A. and Sliwka, D. (2017). Social responsibility and incentives in the lab: Why do agents exert more effort when principals donate? *Journal of Economic Behavior & Organization*, 142:482–493.

Kallgren, C. A., Reno, R. R., and Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and Social Psychology Bulletin*, 26(8):1002–1012.

Klein, N., Grossman, I., Uskul, A. K., Kraus, A., and Epley, N. (2015). It pays to be nice, but not really nice: Asymmetric evaluations of prosociality across seven cultures. *Judgment and Decision Making*, 10:355–364.

Kranton, R. E. (2016). Identity economics 2016: Where do social distinctions and norms come from? *American Economic Review*, 106(5):405–09.

Krupka, E. and Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, 30(3):307–320.

Krupka, E. L. and Weber, R. A. (2013). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *Journal of the European Economic Association*, 11(3):495–524.

Lacetera, N. and Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization*, 76(2):225–237.

Lambarraa, F. and Riener, G. (2015). On the norms of charitable giving in islam: Two field experiments in morocco. *Journal of Economic Behavior & Organization*, 118:69–84.

Lazear, E. P., Malmendier, U., and Weber, R. A. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics*, 4(1):136–63.

Linardi, S. and McConnell, M. A. (2011). No excuses for good behavior: Volunteering and the social environment. *Journal of Public Economics*, 95(5-6):445–454.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493.

Madrian, B. C. (2014). Applying insights from behavioral economics to policy design. *Annu. Rev. Econ.*, 6(1):663–688.

Matland, R. E. and Murray, G. R. (2016). I only have eyes for you: Does implicit social pressure increase voter turnout? *Political Psychology*, 37(4):533–550.

Milgrom, P. R., North, D. C., and Weingast*, B. R. (1990). The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Economics & Politics*, 2(1):1–23.

Nowak, M. A. and Sigmund, K. (1998a). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, 194(4):561–574.

Nowak, M. A. and Sigmund, K. (1998b). Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685):573.

Prat, A. (2005). The wrong kind of transparency. *American economic review*, 95(3):862–877.

Reijula, S., Kuorikoski, J., Ehrig, T., Katsikopoulos, K., Sunder, S., et al. (2018). Nudge, boost, or design? limitations of behaviorally informed policy under social interaction. *Journal of Behavioral Economics for Policy*, 2(1):99–105.

Richter, I., Thøgersen, J., and Klöckner, C. (2018). A social norms intervention going wrong: Boomerang effects from descriptive norms information. *Sustainability*, 10(8):2848.

Rilke, R. M., Danilov, A., Irlenbusch, B., Weisel, O., and Shalvi, S. (2018). The honest leader effect-how hierarchies affect honesty in groups. In *Academy of Management Proceedings*, volume 2018, page 10365. Academy of Management Briarcliff Manor, NY 10510.

Rogers, T., Goldstein, N. J., and Fox, C. R. (2018). social mobilization. *Annual Review of Psychology*, 69:357–381.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

Snowberg, E. and Yariv, L. (2018). Testing the waters: Behavior across participant pools. Working Paper 24781, National Bureau of Economic Research.

Sunstein, C. R. (2017). Nudges that fail. *Behavioural Public Policy*, 1(1):4–25.

Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A., and Aczel, B. (2018). A systematic scoping review of the choice architecture movement: Toward understanding when and why nudges work. *Journal of Behavioral Decision Making*, 31(3):355–366.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

Von Hippel, W. and Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, 34(1):1.

Xiao, E. and Houser, D. (2011). Punish in public. *Journal of Public Economics*, 95(7-8):1006–1017.

Appendix

I. Laboratory Experiment

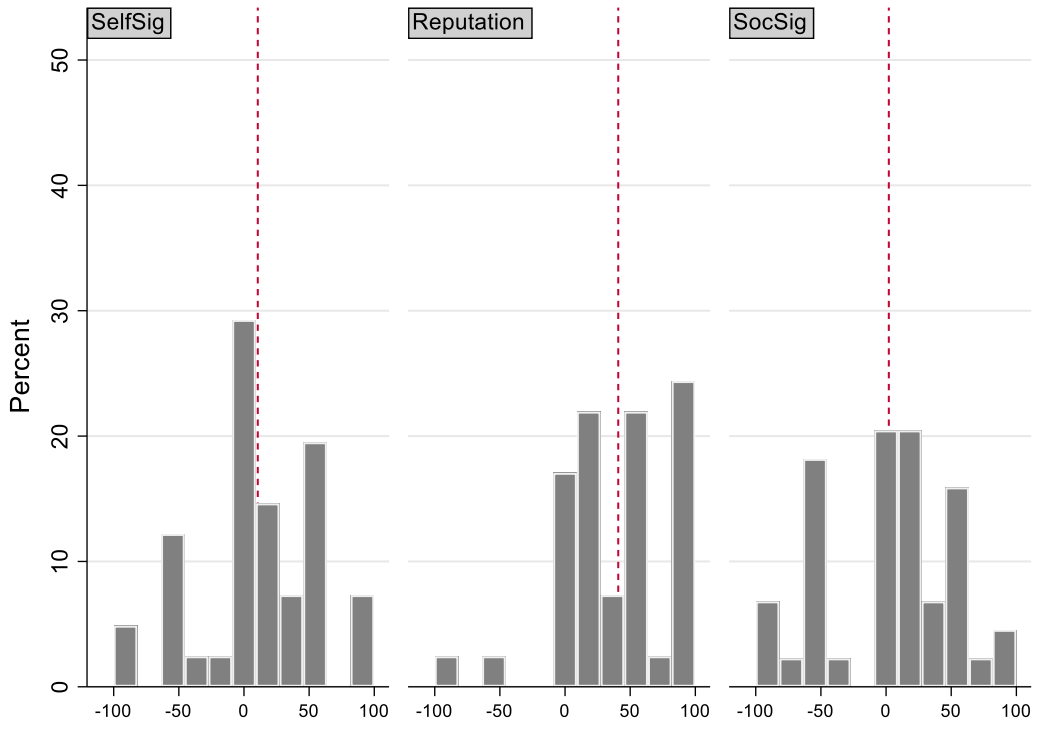

Figure A.1: Distribution of change in charity account (compared to initial endowment) across treatments.

We examine the magnitude of change in the charity account in Figure A.2 and present the regression analysis in Table A.1.

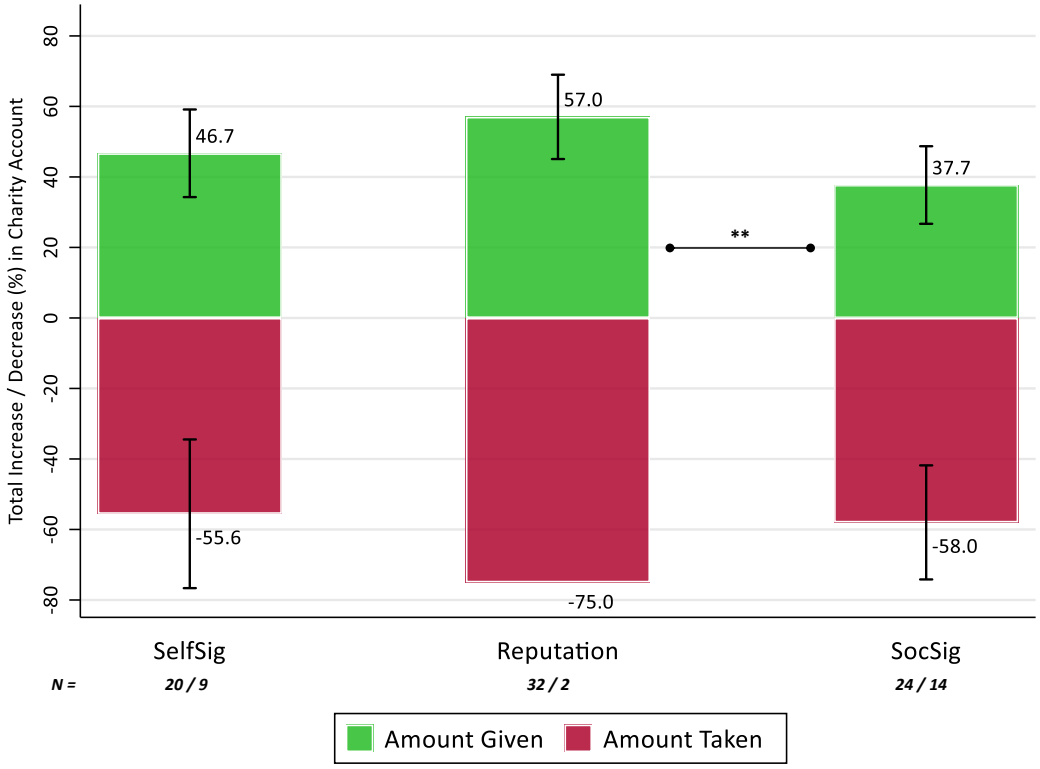

Figure A.2: Magnitude of change in charity account (compared to initial endowment) across treatments. Observations per treatment are displayed at the bottom of each column. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels. Whiskers represent 95% CIs.

Tobit	**Total Change**	**Taking Behavior Only**	**Giving Behavior Only**
DV: Magnitude of Behavior			
Treatment (Base: SelfSig)			
T1 (Reputation)	36.7121***	-7.6487	16.5604
	(11.7116)	(29.1729)	(10.1505)
T2 (SocSig)	-8.5966	-10.1245	-7.6050
	(11.4833)	(15.4497)	(10.2936)
Male	-5.9332	-30.4434*	13.4626
	(10.3051)	(17.0774)	(8.9359)
Belief	0.1302*	-0.0580	0.1204**
	(0.0681)	(0.0948)	(0.0583)
Risk	11.6083*	7.0525	0.7836
	(5.8973)	(7.6509)	(5.0126)
Self-Control	11.2135**	-0.5202	3.8886
	(4.9180)	(7.7866)	(4.2690)
Charity Important	2.5261	2.4012	0.1741
	(2.8858)	(3.3773)	(2.7240)
Charity			
UNICEF	-9.2678	27.2867	-1.3159
	(12.0584)	(18.8187)	(10.9712)
WWF	12.0783	-19.0679	7.9163
	(12.0132)	(21.4336)	(9.7211)
Constant	-23.9032	-58.0120*	24.6867
	(26.5569)	(29.3506)	(27.0542)
Post-estimation test T1 vs T2	p<0.01	p=0.93	p=0.011
Observations	126	25	76

Table A.1: Tobit regressions with standard errors in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively. Results are available upon request. *Gender* (1 = male); *Belief* (amount A beliefs about B's reimbursement decision); *Risk* (higher number = more risk-seeking, standardized measure); *Self-Control* (higher number = more self-control, standardized measure); *Charity Important* (higher number = higher preference for charities in general); *Charity* (higher number = higher preference for the respective charity, Doctors Without Borders is the reference category).

We break out the frequency of behavior across treatments in Figure A.3 and present the analogous regression analysis in Table A2. Focusing first on the shift in the Reputation distribution we observe much as we would expect under Hypotheses 1* and 2* and the magnitude shift analyzed above: Anti-social behavior is significantly less frequent in the Reputation treatment compared to SelfSig (4.88% vs. 21.95%, Equality of Proportions Test (EPT), p=0.02) and SocSig (4.88% vs. 31.82%, EPT, p<0.01). Pro-social behavior is significantly more frequent in the Reputation treatment compared to SelfSig (78.05% vs. 48.72%, EPT, p<0.01) and SocSig (78.05% vs. 54.55%, EPT, p=0.02).

The analogous investigation of 'no change' reveals a weakly significant decrease in SocSig relative to SelfSig (13.6% vs. 29.3%, EPT, p=0.078). Further, we observe two sizable but insignificant changes in behavior: as anticipated by Hypothesis 1*, the directional change of giving behavior indicates an increase relative to SelfSig (54.5% vs. 48.8%, EPT, p=0.595). But contrary to Hypothesis 1*, the directional change of taking behavior yields an increase as well (31.8% vs. 22.0%, EPT, p=0.306). Nevertheless, the offsetting increase, relative to SelfSig, in both giving and taking behavior helps to explain why the the total contribution level fails to rise in SocSig as anticipated by Hypothesis 1*. All shifts between SocSig and Reputation treatments shown in Figure A.3 are consistent with Hypothesis 2*: giving behavior rises and taking behavior falls significantly.[32]

[32] A further analysis relying on Fisher's exact chi-squared statistics reveals that behavior in our *Reputation* treatment differs substantially and significantly from behavior in the *SelfSig* treatment (p=0.02) and *SocSig* treatment (p<0.01). These differences can be attributed to an (under-) over-proportional observation of taking and giving behavior in the reputation treatment. Our chi-squared test statistic yields a p-value of 0.2 when comparing the distribution of behavior in the SelfSig versus SocSig treatments.

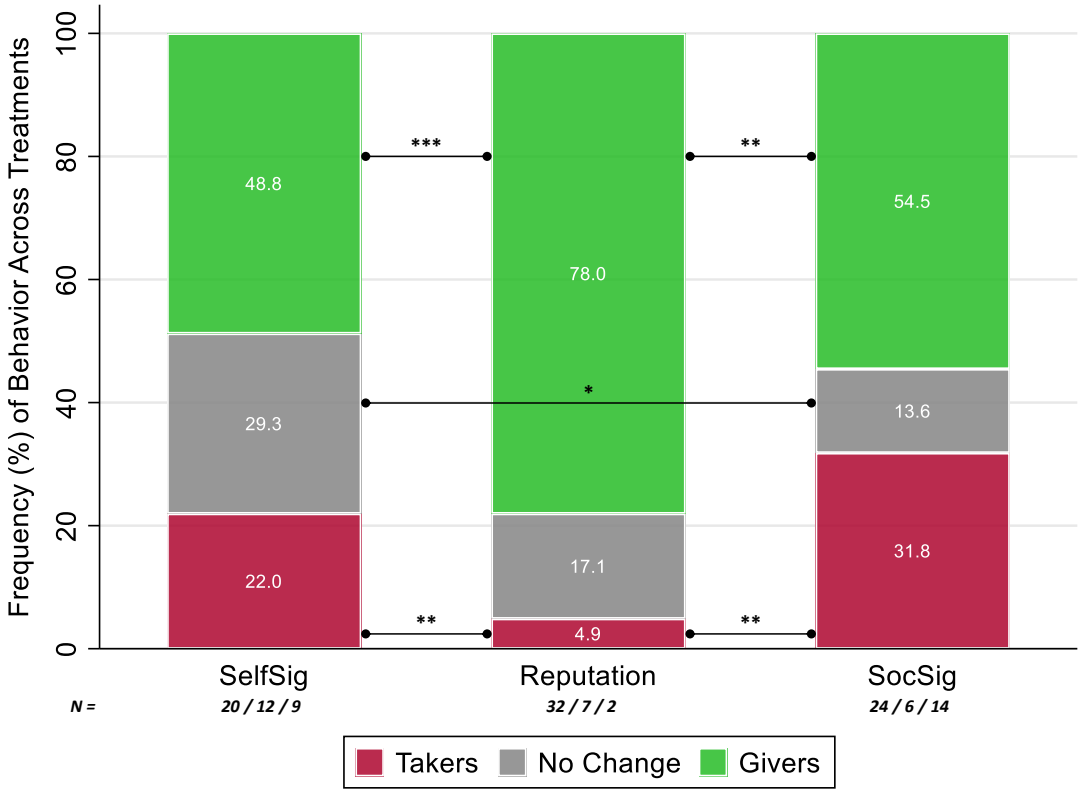

Figure A.3: Frequency of behavior across treatments. Observations per treatment and type of behavior are displayed at the bottom of each column. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels.

Logit regressions reported in Table A.2 use SelfSig as the reference category and, like our previous regression analysis, control for a number of relevant covariates (gender, beliefs, measures of risk and self-control, and importance of charities). The results corroborate the previous non-parametric findings, indicating that reputation concerns, as presented in our Reputation treatment, reduce (increase) the frequency of taking (giving) behavior relative to both self-signaling (SelfSig) and social signaling (SocSig) concerns.

The coefficients of the covariates indicate that males engage more frequently in anti-social behavior, which is in line with findings on corruption and cheating behavior in existing literature (for a discussion, see Croson and Gneezy, 2009; Dimant and Tosato, 2018; Abeler et al., 2018). Conversely, we do not observe a statistically significant gender

heterogeneity with respect to pro-social behavior. Our results also indicate that Player A's beliefs, the extent to which she thinks she will be reimbursed by Player B's transfer from the cash box, only weakly predict giving behavior. Notably, while the direction of the coefficients for SocSig suggest that both giving and taking behavior are more frequent compared to SelfSig, they do not reach significance.

Logit	**Taking Behavior**				**Giving Behavior**			
DV: Frequency of Behavior	(1a) Odds Ratios	(1b) Log-Odds	(2a) Odds Ratios	(2b) Log-Odds	(1a) Odds Ratios	(1b) Log-Odds	(2a) Odds Ratios	(2b) Log-Odds
Treatment (Base: Anonymous)								
T1 (Reputation)	0.1823**	-1.7019**	0.1403**	-1.9637**	3.7333***	1.3173***	4.1647***	1.4266***
	(0.1496)	(0.8206)	(0.1315)	(0.9367)	(1.8362)	(0.4918)	(2.2259)	(0.5345)
T2 (SocSig)	1.6593	0.5064	1.9133	0.6488	1.2600	0.2311	1.4302	0.3578
	(0.8281)	(0.4991)	(1.0599)	(0.5540)	(0.5504)	(0.4368)	(0.6629)	(0.4635)
Gender			3.3503**	1.2090**			1.1165	0.1102
			(1.9703)	(0.5881)			(0.5079)	(0.4549)
Belief			1.0010	0.0010			1.0052*	0.0052*
			(0.0039)	(0.0039)			(0.0031)	(0.0031)
Risk			0.4718**	-0.7512**			1.2776	0.2450
			(0.1707)	(0.3619)			(0.3159)	(0.2472)
Self-Control			0.5561**	-0.5867**			1.4150*	0.3471*
			(0.1573)	(0.2829)			(0.2951)	(0.2086)
Charity Important			0.9501	-0.0512			1.1969	0.1798
			(0.1466)	(0.1543)			(0.1454)	(0.1215)
Charity								
UNICEF			1.4981	0.4042			0.4445*	-0.8107*
			(0.8601)	(0.5741)			(0.2131)	(0.4794)
WWF			0.5777	-0.5486			1.3952	0.3330
			(0.4460)	(0.7720)			(0.7452)	(0.5341)
Post-estimation tests: T1 vs. T2	p<0.01		p<0.01		p=0.0253		p=0.0404	
Observations	126	126	126	126	126	126	126	126

Table A.2: Logistic Regression. Coefficients denote Odds Ratios (OR). Standard errors in parenthesis. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively. *Gender* (1 = male); *Belief* (amount A beliefs about B's reimbursement decision); *Risk* (higher number = more risk-seeking, standardized measure); *Self-Control* (higher number = more self-control, standardized measure); *Charity Important* (higher number = higher preference for charities in general); *Charity* (higher number = higher preference for the respective charity, Doctors Without Borders are the reference category).

Experimental Design

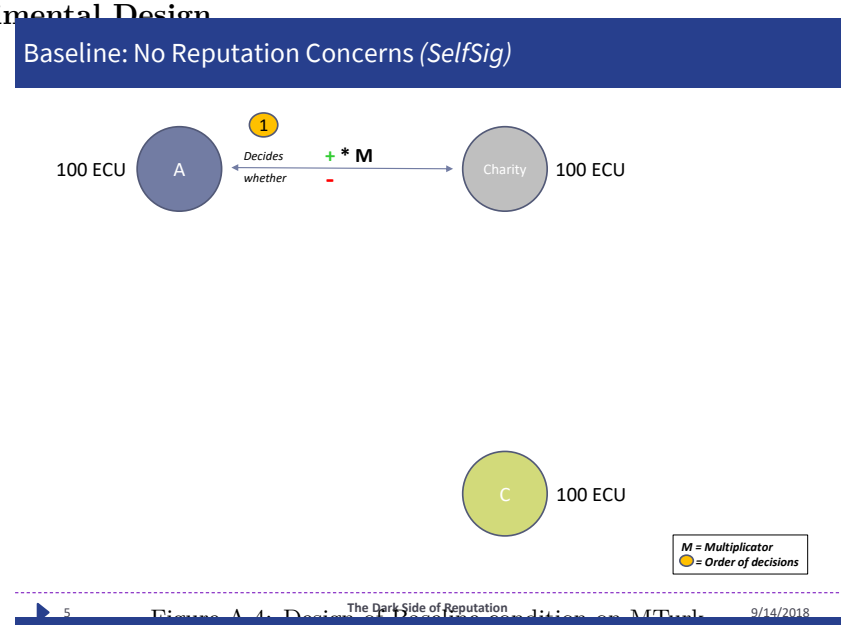

Figure A.4: Design of Baseline condition on MTurk.

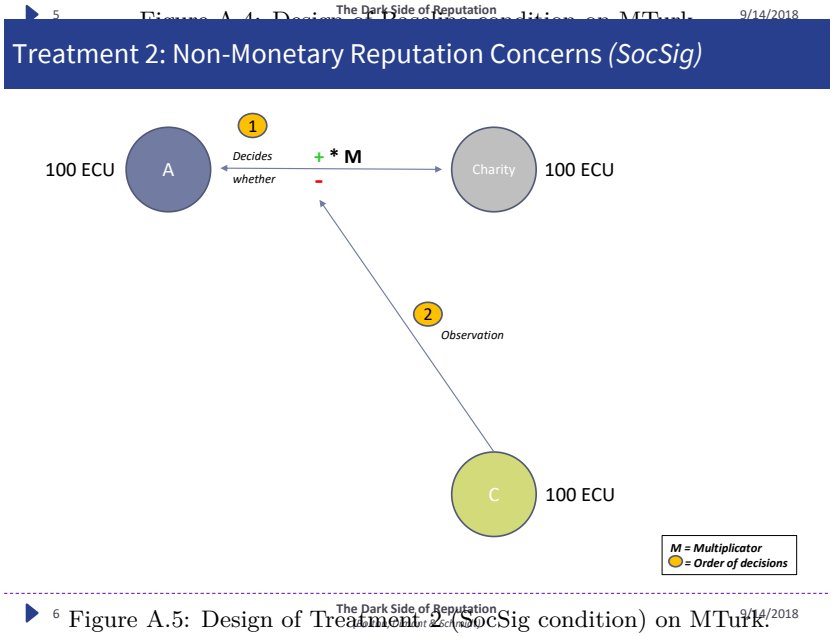

Figure A.5: Design of Treatment 2 (SocSig condition) on MTurk.

II.a. MTurk Experiment 1: Role of Inequality Concerns

The regression analyses below correspond to the graphical illustration in Figure 8.

Tobit	**Total Change**	**Taking Behavior Only**	**Giving Behavior Only**
DV: Magnitude of Behavior			
SocSig	13.7226**	-28.5246**	-3.2735
	(6.3549)	(13.0066)	(8.4866)
Disadvantageous Inequality (DI)	1.5418	27.2151**	-28.5516***
	(6.6806)	(11.2398)	(9.1925)
SocSig × Disadvantageous Inequality (DI)	-29.1155***	25.6800	-14.4037
	(9.2616)	(16.6277)	(12.6173)
Male	1.6932	6.7180	-10.6302*
	(4.6811)	(7.9960)	(6.4207)
Risk	0.0615	2.5901	-0.4161
	(0.9500)	(1.5821)	(1.2360)
Age	-0.2404	-0.0054	-0.3048
	(0.2170)	(0.3695)	(0.3134)
Constant	-15.8804	-152.3835***	108.5449***
	(13.7399)	(23.8957)	(19.0318)
Observations	831	376	209

Table A.3: Tobit Regression. Standard errors in parenthesis. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively. *Gender* (1 = male); *Risk* (higher number = more risk-seeking, standardized measure); *Age* (higher number = older).

We examine the frequency of behavior change in Figure A.6 and present the regression analysis in Table A.4.

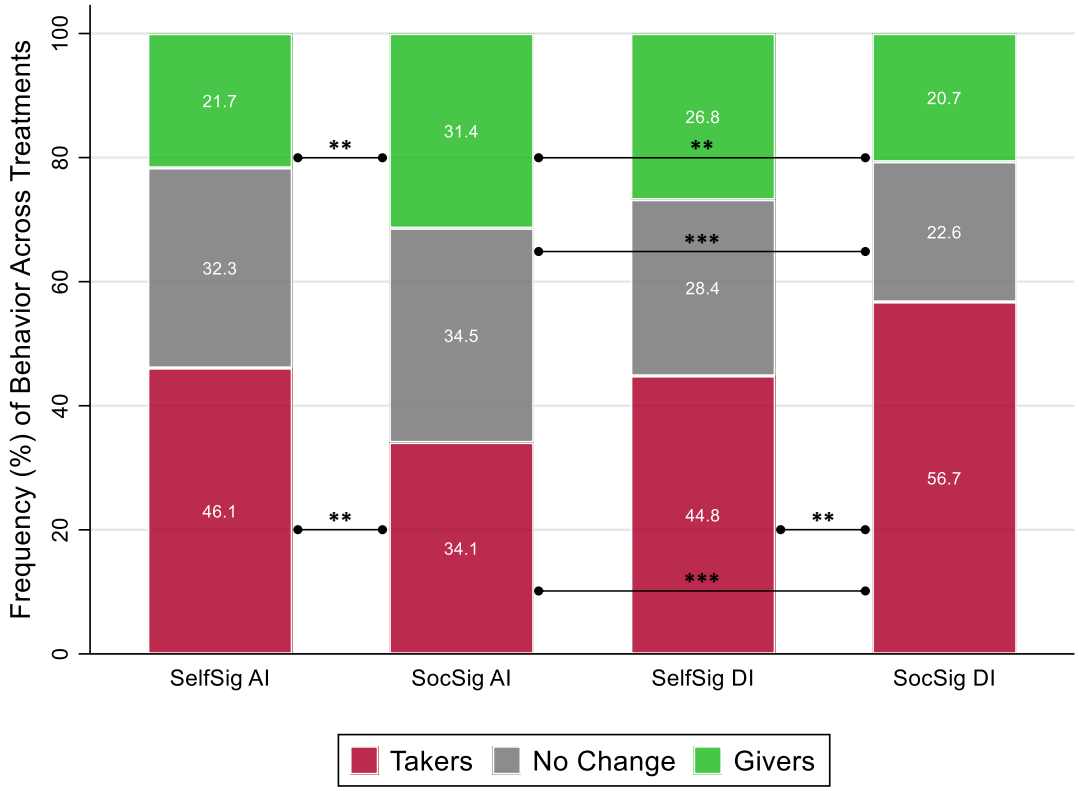

Figure A.6: Change in charity account (in frequencies) across treatments. Observations per treatment are displayed at the bottom of each column in the top-down order. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels.

Logit	**Taking Behavior**		**Giving Behavior**	
DV: Frequency of Behavior	(1a) Odds Ratios	(1b) Log-Odds	(2a) Odds Ratios	(2b) Log-Odds
SocSig	0.6061** (0.1192)	-0.5008** (0.1967)	1.6621** (0.3677)	0.5081** (0.2212)
Disadvantageous Inequality (DI)	0.9496 (0.1926)	-0.0517 (0.2028)	1.3079 (0.3096)	0.2684 (0.2367)
SocSig × Disadvantageous Inequality (DI)	2.6811*** (0.7609)	0.9862*** (0.2838)	0.4364** (0.1422)	-0.8292** (0.3258)
Male	0.8563 (0.1230)	-0.1551 (0.1436)	1.2068 (0.1969)	0.1880 (0.1632)
Risk	0.9946 (0.0289)	-0.0054 (0.0291)	0.9529 (0.0317)	-0.0482 (0.0332)
Age	1.0054 (0.0067)	0.0054 (0.0067)	0.9940 (0.0074)	-0.0060 (0.0074)
Observations	831	831	831	831

Table A.4: Logistic Regression, odds ratios displayed. Coefficients denote Odds Ratios (OR). Standard errors in parenthesis. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively. *Gender* (1 = male); *Risk* (higher number = more risk-seeking, standardized measure); *Age* (higher number = older).

III.b. MTurk Experiment 2: Role of Norm Focus

Figure A.7 illustrates the stated beliefs in our norm focus intervention treatments.

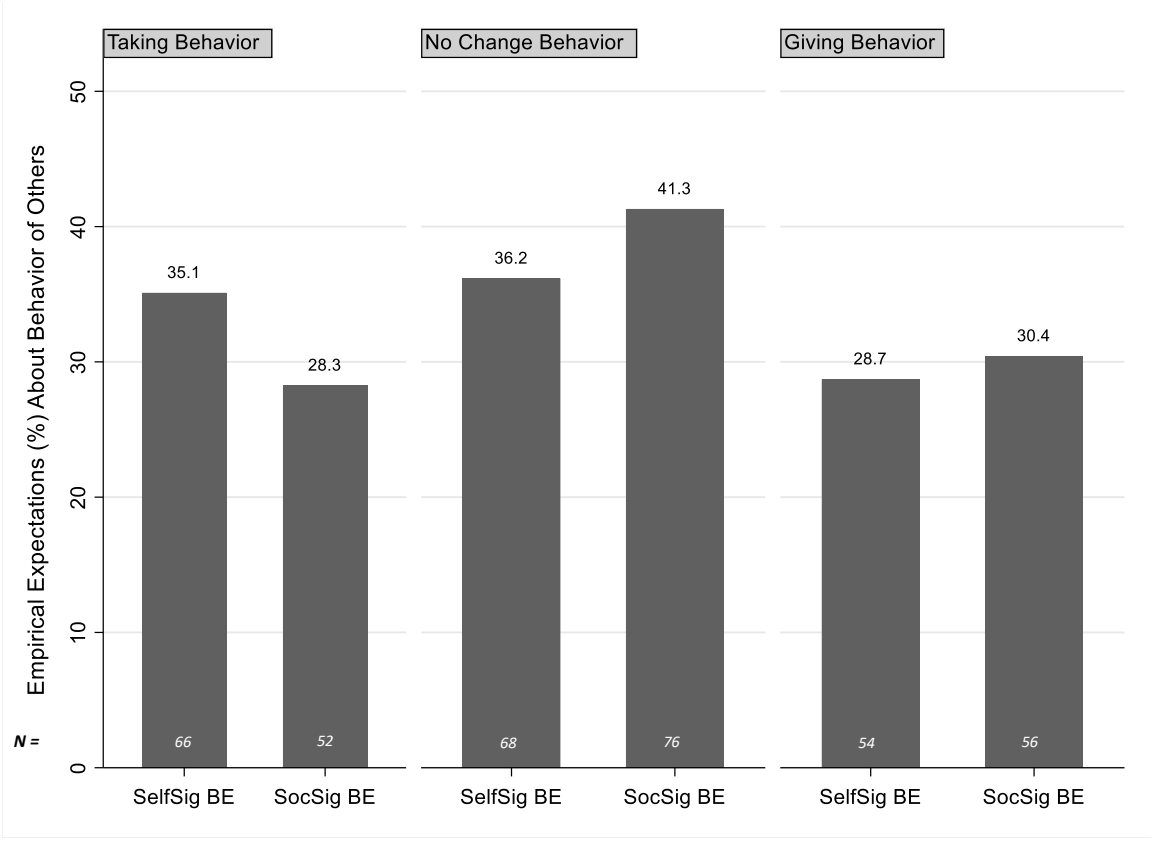

Figure A.7: Beliefs in SelfSig BE and SocSig BE conditions. None of the differences reaches significance. Taking behavior: MWU, p = 0.156; No Change Behavior: MWU, p = 0.309; Giving Behavior: MWU, p = 0.718.

The data suggests that Result 1 – more extreme behavior in SocSig compared to SelfSig – is explained through both more extreme giving behavior, (66.4%, vs. 51.7%, MWU, p=0.018) and taking behavior (-90.6% vs. -81.4%, MWU, p=0.012) alike. For giving behavior alone, we find that the norm focus improves behavior for SelfSig (65.4% vs. 51.7%, MWU, p=0.018) but does not at all affect giving in SocSig. For taking behavior alone, our results suggest that norm focus significantly reduces the extent of taking for SocSig (-72.6% vs. -90.6%, MWU, p<0.001), while the norm focus intervention has no effect on the SelfSig setup.

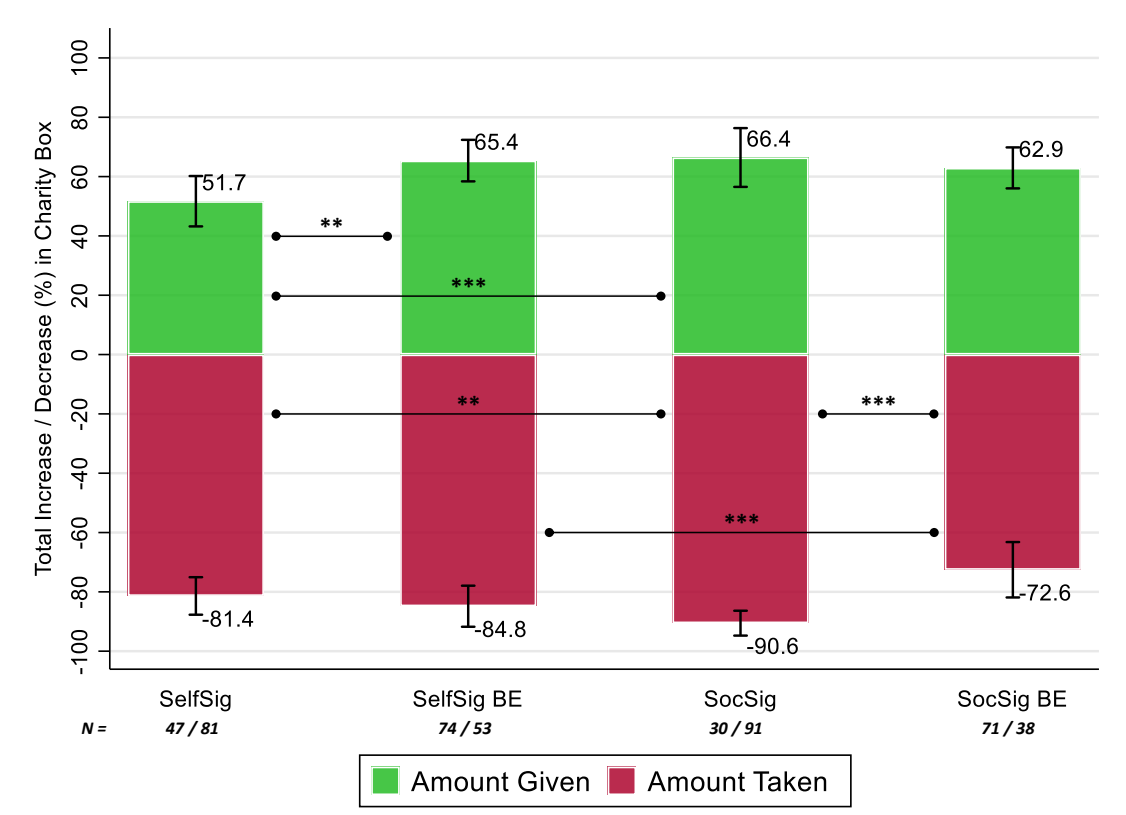

Figure A.8: Change in charity account (compared to initial endowment) across treatments conditional on type of behavior. Observations per treatment and type of behavior are displayed at the bottom of each column in the top-down order. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels. Vertical lines represent 95% CIs.

We ensure the robustness of these results through the lens of Tobit regressions (to account for the censoring nature of the data) and present the results in Table A.5. While this

analysis corroborates the main non-parametric findings discussed above, we also observe a number of additional noteworthy insights. For one, we observe that norm focus yields an over-proportional increase in the charity account for SocSig. When subdividing this total effect into its sub-components, we see that this result is driven by an extreme and highly significant upward reaction of taking behavior and an equally significant but comparably smaller downward reaction for giving behavior. The positive (negative) coefficient for taking (giving) behavior indicates that amplifying social signaling with a norm focus leads to a substantial reduction of this behavior relative to behavior in the SelfSig conditions.

Tobit	**Total Change**	**Taking Behavior Only**	**Giving Behavior Only**
DV: Magnitude of Behavior			
SocSig	-15.2400**	-31.4818***	18.2704*
	(7.7702)	(11.3326)	(9.6754)
Norm Focus (BE)	30.4678***	-8.1062	18.1138**
	(8.3484)	(12.3039)	(7.8773)
SocSig × Norm Focus (BE)	24.1741**	57.7633***	-24.0088**
	(11.4463)	(18.0168)	(11.8613)
Male	18.5412***	17.8013*	7.5523
	(5.8738)	(9.1836)	(5.6006)
Risk	-1.3118	-0.1500	1.1333
	(1.2378)	(1.7779)	(1.2934)
Age	0.4250	0.3174	0.2830
	(0.2787)	(0.4501)	(0.2863)
Constant	-63.0140***	-146.5984***	29.3511*
	(16.3344)	(24.2465)	(16.8338)
Observations	819	263	222

Table A.5: Tobit Regression. Standard errors in parenthesis. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively. *Gender* (1 = male); *Risk* (higher number = more risk-seeking, standardized measure); *Age* (higher number = older).

We present the distribution of behavior across treatments in Figure A.9.

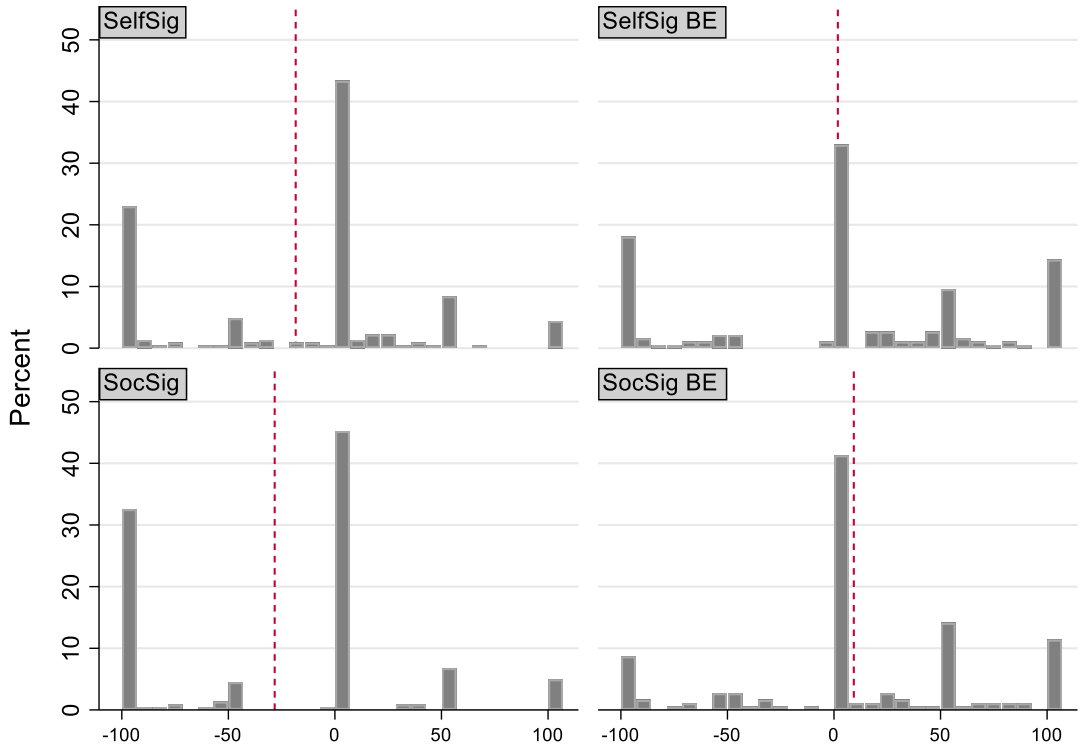

Figure A.9: Distribution of change in charity account (compared to initial endowment) across treatments.

A noteworthy finding is that the initial surprising result from Experiment 1 reproduces, and now yields, a statistically significant difference between SelfSig and SocSig in the form of *reduced* giving behavior frequency in SocSig (13.6% vs. 20.8%, EPT, p=0.042).[33] We do not observe any significant differences for taking or no change behavior. We present the results in Figure A.10.

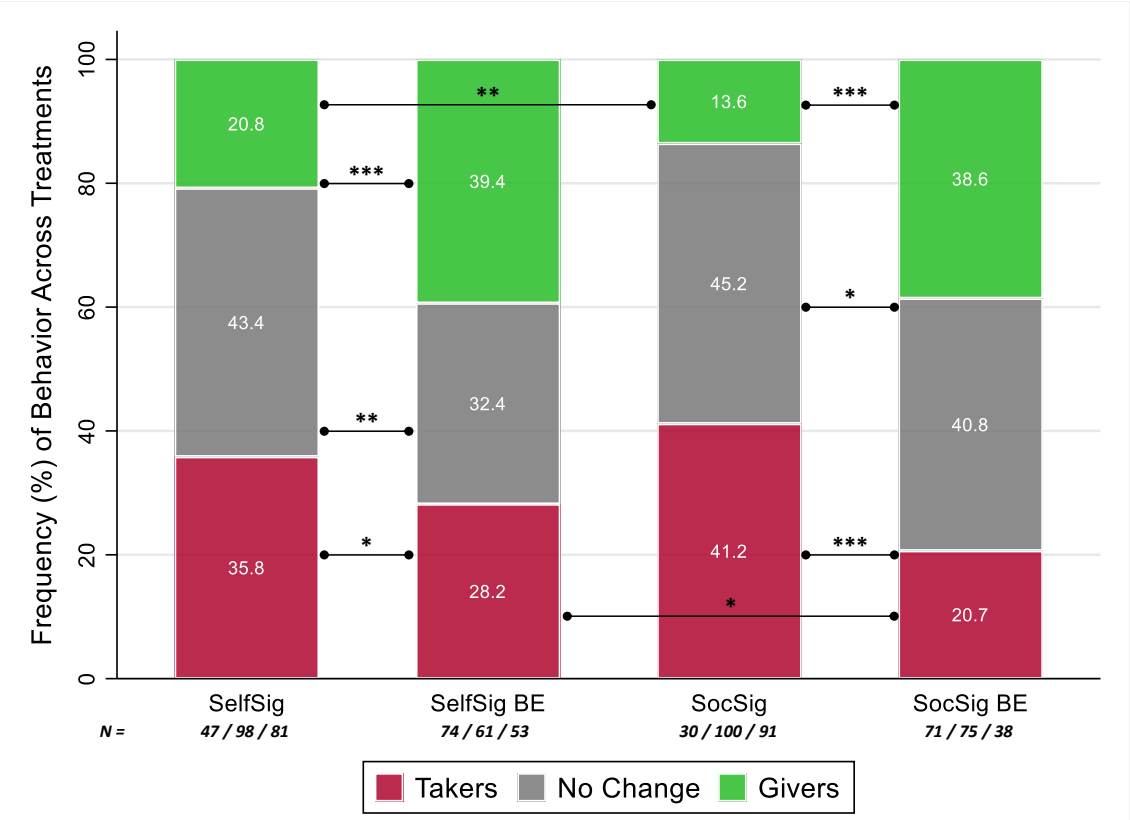

Figure A.10: Change in charity account (in frequencies) across treatments. Observations per treatment are displayed at the bottom of each column in the top-down order. Horizontal lines with stars represent statistical significance at *** p<0.01, ** p<0.05, and * p<0.1. Absence of horizontal implies lack of statistical significance at the conventional levels.

The results also indicate that norm focus has the largest impact on the frequency of giving behavior, both in the SelfSig (39.4% vs. 20.8%, EPT, p<0.001) and SocSig (36.8%

[33]This finding can be harmonized not only with higher statistical power, but also with respect to the Focus Theory: the absence of Player B leads to a reduced focus, in particular in the SocSig condition, allowing for the difference with SelfSig to be more strongly pronounced than in Experiment 1.

vs. 13.6%, EPT, p<0.001) conditions. This stark increase in compliance comes at the expense of both taking and no change behavior. In particular, we observe that an increased norm focus also yields significantly less frequent taking behavior, again both for the SelfSig (28.2% vs. 35.8%, EPT, marginally significant at p=0.098) as well as the SocSig (20.7% vs. 41.2%, EPT, p<0.001) conditions. Lastly, we observe that the focus intervention also leads to a significant reduction of no change behavior (32.4% vs. 43.4%, EPT, p=0.023; 40.8% vs. 45.2%, p=0.096).

We perform a Logit regression (Table A.6) to tease out the extent to which both social signaling as well as the norm focus affect the likelihood of individuals to engage in either taking or giving behavior after controlling for gender, age, and risk attitudes to capture specific characteristics of MTurkers. In line with our non-parametric results presented above, social signaling only reduces (marginally) the frequency of giving behavior, while the effect of norm focus in SelfSig has a substantial effect in reducing taking and increasing giving behavior. We also find indications of a marginally significant and negative interaction effect for taking behavior, suggesting that a social signaling nudge can be rendered more effective in reducing the frequency of anti-social behavior when paired with a norm focus. The absence of a significant interaction for giving behavior suggests that pairing a social signaling nudge with a norm focus intervention can eliminate the backfiring effect.

Logit	**Taking Behavior**		**Giving Behavior**	
DV: Frequency of Behavior	(1a) Odds Ratios	(1b) Log-Odds	(2a) Odds Ratios	(2b) Log-Odds
SocSig	1.2590 (0.2458)	0.2303 (0.1952)	0.5965** (0.1525)	-0.5168** (0.2557)
Norm Focus (BE)	0.6148** (0.1382)	-0.4865** (0.2249)	2.3830*** (0.5525)	0.8683*** (0.2319)
SocSig × Norm Focus (BE)	0.5403* (0.1716)	-0.6157* (0.3175)	1.6086 (0.5363)	0.4753 (0.3334)
Male	0.6246*** (0.0987)	-0.4706*** (0.1580)	1.0708 (0.1809)	0.0684 (0.1689)
Risk	1.0658* (0.0359)	0.0637* (0.0337)	1.0184 (0.0353)	0.0182 (0.0347)
Age	0.9881 (0.0076)	-0.0119 (0.0077)	0.9985 (0.0085)	-0.0015 (0.0085)
Observations	819	819	819	819

Table A.6: Logistic Regression, odds ratios displayed. Coefficients denote Odds Ratios (OR). Standard errors in parenthesis. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively. *Gender* (1 = male); *Risk* (higher number = more risk-seeking, standardized measure); *Age* (higher number = older).

III.c. Follow-Up Norm Elicitation on MTurk

We use the method by Krupka and Weber (2013) to elicit incentive-compatible normative beliefs about the appropriateness of behavior in the context of our setup. For this, we collect data from 200 MTurk participants, none of which have previously played any version of our game. The results are conclusive: giving is the norm.

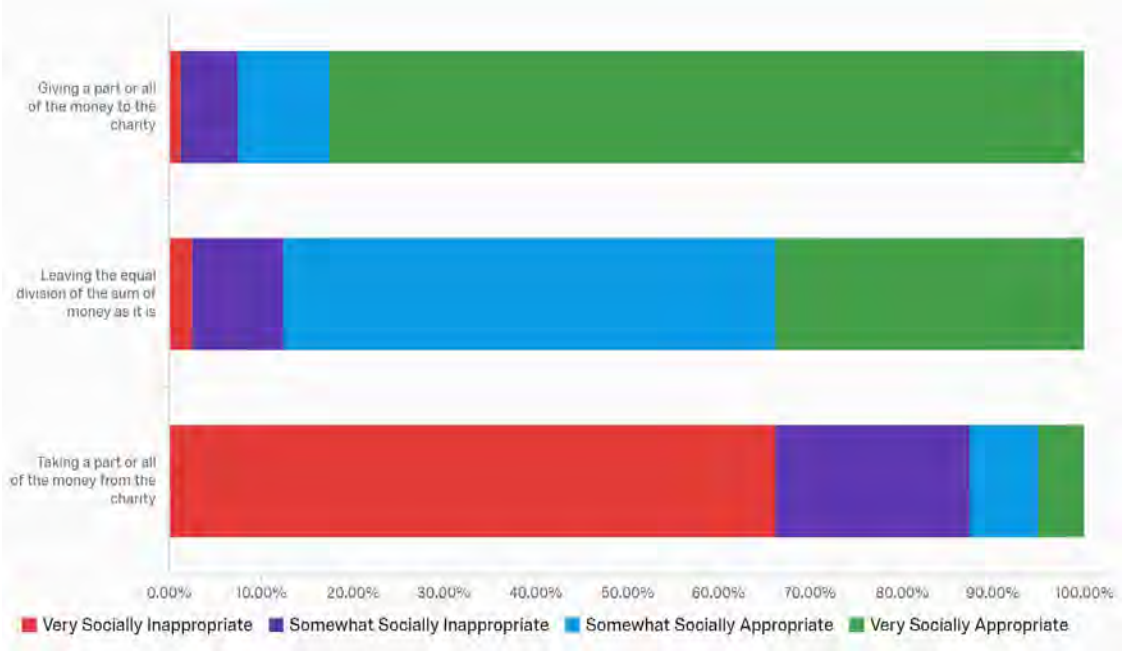

Figure A.11: Belief-elicitation method following Krupka and Weber (2013).

III. Instructions

In what follows, we present the instructions used in the laboratory experiment. The instructions of the follow-up experiment on MTurk followed the same logic and were adjusted appropriately. Details are available upon request from the authors.

Instructions

- First of all, we would like to thank you very much for participating in this experiment. Please read the instructions carefully. The experiment will last for about 45-60 minutes.
- During the entire experiment, no communication is allowed. If there is something that you do not understand or if you have any questions, now or at some point during the experiment, please raise your hand and remain seated. One of our colleagues will come to you and answer your question.
- During the experiment, you have the possibility to earn money. The amount you will receive at the end of the session depends on how many "Taler" you earn during the experiment.
- At the end of the experiment, the amount of "Taler" that you have earned will be converted into real money at an exchange rate of 10 Taler = 1 Euro.
- You will receive a show-up fee of 100 Taler with which you will play throughout the experiment. Depending on your role and your decisions, you will be able to make additional money. You will be paid anonymously at the end of the experiment.

Procedure

- The experiment consists of a questionnaire and decisions. The experiment will be played exactly once.
- There are 3 different types of roles in this experiment: **A**, **B**, and **C**. Player A moves first, after which Players B and C follow. In addition to this, a charity will be part of this experiment with which only Player A interacts.
- Each participant in this experiment is randomly assigned one of these three roles at the beginning of the experiment and remains in the same role throughout the experiment. No participant plays in more than one role.
- Each participant plays this experiment in groups of 3, each of which are in a different role. Hence, each group has exactly one participant in the role of A, B, and C, respectively. Each participant as well as the charity start with the exact same endowment of 100 Taler.
- The decision of Player A only affect the charity. The exact charitable organization will be randomly chosen from a list of 3 charities at the end of the experiment.
- The decision of Player B only affects the payoffs of Player A.
- Player C does not make any active decisions.

Player A's Role in Detail

- A player in this role starts with his or her show-up fee of 100 Taler. The charitable organization with which this player interacts starts with 100 Taler as well.
- Player A can make one of the following decisions towards the charity exactly once:
    1. To take money away from the charity and add to one's own account
    2. To leave the equal split between him/her and the charity unaltered
    3. To give money from one's own account to the charity (any amount given to charity will be doubled by the experimenter.
- The amount of money that is in the charity account at the end of the experiment will be doubled by the experimenter and donated to one of the following three charities: (1) Doctors Without Borders, (2) United Nations Children's Fund (UNICEF), or (3) World Wide Fund for Nature (WWF). The charity will be randomly chosen at the end of the experiment.
    - To ensure the credibility of our claim, we will upload all donation receipts to our website in due time.
- Player A's decision towards the charity may or may not be observed by either Player B or Player C. This will be randomly determined once the experiment starts and announced on the screen.

Player B's Role in Detail

- A player in this role starts with his or her show-up fee of 100 Taler.
- In addition, this player is given the right to decide upon the distribution of an additional 100 Taler in a *money box*.
- This player decides whether and how much of the 100 Taler in this *money box* he/she wants to send to Player A. The mount sent will be doubled by the experimenter and added to Player's A account.
- Importantly, the money left in this *money box*, i.e. money that Player B decided not to send to Player A, cannot be retained by Player B. That is, Player B retains his/her original endowment of 100 Taler regardless of his/her decision of what to do with the 100 Taler in the *money box*.

Player C's Role in Detail

- A player in this role starts with his or her show-up fee of 100 Taler.
- Player C remains passive and does not engage in any active decision-making.

General Payoff Procedure

- At the end of the experiment, exactly one group in this session that you partake is chosen at random. Only the decisions of the members of this group are implemented. Based on this, both Player A and B, whose decisions are payoff-relevant, will be players from the same group.
    - This means that the decision of exactly one Player A in this experiment is going to determine the payoffs of the charitable organization.
    - This also means that the decision of exactly one Player B will affect exactly one Player A's payoff.
    - In this case, Player A can leave with a minimum of 0 Taler (if he/she decides to give all of the initial endowment to the charity) and a maximum of 400 Taler (if he/she decides to keep the own initial endowment (= 100 Taler), take all of the money from the charity (100 Taler), and receives all of the money from the charity by Player B (= 2*100 Taler)).
- Every other participant who is not part of the randomly chosen group will receive his/her initial endowment of 100 Taler regardless of his/her decisions. In those groups, no decisions towards the charity are payoff relevant. The charity will only be paid once by the group that is payoff relevant.
- Whether or not you are part of the payoff-relevant group will be revealed only at the very end of the experiment.

Detailed Payoff Procedure

If player ends up in the randomly chosen payoff-relevant group:

- Player A: (100 Taler initial endowment) **+/-** (Taler given to/taken from the charity) **+** (Taler received from Player B)
- Player B: 100 Taler
- Player C: 100 Taler
- Charity:
    - If Player A decides to give money: 100 Taler + 2*amount given by Player A
    - If Player A decides to take money away: 100 Taler – amount taken by Player A

If player does not end up in the randomly chosen payoff-relevant group:

- Player A: 100 Taler
- Player B: 100 Taler
- Player C: 100 Taler
- Charity: 0 (decisions towards the charity not payoff relevant)