Discussion Papers of the
Max Planck Institute for
Research on Collective Goods
Bonn 2018/16

The differential effect of narratives

Adrian Hillenbrand
Eugenio Verrina

MAX PLANCK SOCIETY

Discussion Papers of the
Max Planck Institute
for Research on Collective Goods

Bonn 2018/16

The differential effect of narratives

Adrian Hillenbrand / Eugenio Verrina

December 2018

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

The differential effect of narratives

Adrian Hillenbrand[*] and Eugenio Verrina[†]

December 14, 2018

Abstract

Narratives pervade almost any aspect of our life and play a particularly important role in moral and prosocial decision-making. We study how positive (stories in favor of a prosocial action) and negative (stories in favor of a selfish action) narratives influence prosocial behavior. Our main findings are that positive narratives increase giving substantially, especially for selfish types, compared to a baseline with no narratives. Negative narratives, on the other hand, have a differential effect. Prosocial types decrease their giving, while selfish types give more than in the baseline. We also find that positive narratives lead to a binary response (comply or not comply), while negative narratives induce a more gradual trade-off.

Keywords: Prosocial behavior, narratives, justifications, motivated moral reasoning, dictator game, SVO
JEL Classification: C91, D63, D64, D83, D91

[*] Max Planck Institute for Research on Collective Goods (Bonn). E-mail: hillenbrand@coll.mpg.de.

[†] Max Planck Institute for Research on Collective Goods (Bonn), and University of Cologne. E-mail: verrina@coll.mpg.de.

1 Introduction

Imagine that for some days you have seen a beggar on your way to work. While passing by today, you reach into your pocket to get some change. Imagine that, while doing so, you remember what a colleague told you while passing by the day before. He stated that most of these people are not really needy, but have simply chosen to live soaking up money from people who work hard. Besides, the beggar will spend all the money you give him on alcohol and drugs. He deserves no consideration at all. Now imagine, instead, your colleague told you that he had heard of many heart-breaking stories of people who lost everything and had to live on the street from one day to the other. He continued arguing that the government does not do enough for people in need. We should show some humanity and fight against the unfairness of this wicked capitalistic system. Will you give something to the beggar after recalling one of these two stories? Will you give him more or less than what you had picked from your pocket in the beginning? Will you react differently based on your first tendency to give or not to give something?

Theoretical accounts of motivated moral reasoning (Ditto et al., 2009) emphasize people's deep need to justify their moral behavior not only to others, but especially to themselves. From a fully rational standpoint, these justifications could reflect pieces of evidence an individual uses to inform her choice. However, psychological theories of cognitive dissonance (Festinger, 1962) indicate that such reasons are often used beyond that to resolve tensions between beliefs and actions (Akerlof and Dickens, 1982).[1] In our opening illustration, the tension between a self-interested and a prosocial option can be resolved differently, depending on the story one is told or recalls. We will call these rationales or justifications *narratives*.[2] Our definition is borrowed from Bénabou et al. (2018), who develop a model where individuals with self and social image concerns produce and consume narratives as signals complementing their actions.[3] Narratives accompany

[1] Epley and Gilovich (2016) make a very similar point in their discussion of the mechanics behind motivated reasoning in general.

[2] Bruner (1991) describes how narratives help people to construct their own account of the world. We focus on narratives in the context of prosocial behavior.

[3] Foerster and van der Weele (2018) work out a similar model where two agents with social image concerns can exchange signals about the social returns to an investment in

nearly all our decisions, often playing a decisive role in shaping them. They help explaining economic fluctuations (Shiller, 2017) and broader historical phenomena (Akerlof and Snower, 2016). While narratives are deeply grounded in psychological theories (Bruner, 1991; McAdams, 1988), their relevance in moral and prosocial decision-making has only recently received attention in economics (Bénabou et al., 2018; Foerster and van der Weele, 2018). There is, however, little empirical evidence on how narratives directly affect moral and prosocial behavior.

In this paper, we fill part of this gap by studying how people's prosocial behavior is influenced by narratives. In particular, we look at the impact of positive and negative narratives, as defined by Bénabou et al. (2018), on more selfish or more prosocial people. Positive narratives are arguments endorsing moral or prosocial behavior, e.g, by highlighting the presence of a norm or potential reasons supporting it. Negative narratives, on the other hand, justify immoral or selfish behavior and can operate through various mechanisms; they can, e.g., downplay the negative externalities of an action or alter the normative expectations pending on the decision-maker.[4]

In our experiment, subjects play a dictator game where they decide how to share a given amount of money with another anonymous participant. In our two treatment conditions, they are shown either negative or positive narratives before making their choice. Narratives in the NEGATIVE condition are arguments in favor of the selfish action, i.e., giving zero to the other participant, while narratives in the POSITIVE condition are reasons in favor of the prosocial action, i.e., the equal split.[5] Narratives in both treatments arise endogenously from our design as the most convincing arguments used by subjects to justify their choice. We compare these treatments to a BASELINE condition with no narratives. Importantly, we keep

a public good in a simultaneous pre-play communication phase. Their model generates a set of predictions for the use of the signals that are comparable with Bénabou et al. (2018) for what concerns the focus of this paper.

[4] We focus on prosocial behavior as an important component of moral behavior. As opposed to prosocial behavior, we equate immoral behavior to selfish behavior.

[5] Krupka and Weber (2013) provide empirical evidence that this is indeed believed to be the most socially appropriate behavior in the dictator game. In this sense, what we label as the prosocial action would correspond to the social norm in this context, while what we call the selfish action would be the strongest possible deviation from the social norm. As hinted in the Hypotheses section, our theoretical considerations and main intuitions also hold in the context of a normative model.

empirical expectations across all our conditions constant by showing a fixed distribution of choices made in similar experiments. This ensures that our treatment manipulations do not carry any valuable empirical information. We thus isolate the effect of narratives as providing or highlighting reasons for either the selfish or the prosocial action. A key feature of our design is that it allows us to explore how heterogeneous prosocial concerns interact with positive and negative narratives by using subjects' Social Value Orientation (SVO). We also develop a theoretical framework of how giving is influenced by externally supplied narratives and derive simple hypotheses providing a benchmark comparison for our experimental results. According to our model, positive narratives should increase aggregate giving, while negative narratives should decrease it. This effect should go in the same direction for all social types and should be stronger for types close to the extremes.

We find that positive narratives increase giving, while there is no effect of negative narratives at an aggregate level. The latter result is due to a *differential effect* of narratives on different social types. In line with our predictions, types across the whole spectrum increase their giving in the Positive condition, with selfish types displaying the largest effect. However, in the Negative condition prosocial types decrease their giving, while selfish types increase their giving. This result is not in line with our model, in which the same narrative cannot cause effects going in opposite directions for different social types. We offer two potential explanations for this effect: one based on the enhanced salience of the moral decision and another based on a social comparison argument. We also investigate whether narratives lead subjects to fully comply with the behavior they prescribe or just induce them to change their behavior on the margin. We find that the Negative condition does not lead subjects either to follow the narrative or not, i.e., to either split equally or keep everything, more often than in the Baseline. On the other hand, the Positive condition leads more selfish types to split the money equally with a higher probability. This means that, in our setting, negative narratives induce a more gradual trade-off, while positive ones push subjects to fully comply with their prescribed behavior.

2 Related literature

Our work resonates with the growing interest in the role played by narratives (Bénabou et al., 2018; Foerster and van der Weele, 2018; Shiller, 2017; Akerlof and Snower, 2016) and, more generally, the role that motivated reasoning plays in shaping economic interactions (Karlsson et al., 2004; Epley and Gilovich, 2016; Bénabou and Tirole, 2016; Golman et al., 2016; Gino et al., 2016; Carlson et al., 2018; Saucet and Villeval, 2018). Our work is also closely linked to experimental studies on phenomena such as moral wiggle room (Dana et al., 2007; Larson and Capra, 2009; Matthey and Regner, 2011; van der Weele et al., 2014; Feiler, 2014) and to the wider literature investigating self-serving judgments of fairness or morality (Konow, 2000; Hamman et al., 2010; Shalvi et al., 2011a; Wiltermuth, 2011; Rodriguez-Lara and Moreno-Garrido, 2012; Bicchieri and Mercier, 2013; Gino et al., 2013; Shalvi et al., 2015) and self-serving beliefs (Haisley and Weber, 2010; Chance et al., 2011). The main result one can draw from this huge body of evidence is that prosocial behavior is highly sensitive to the specific context in which choices take place, and that people often tweak the evidence in their favor in conscious and unconscious ways. Our work contributes to this growing literature by providing new evidence on how people react to externally provided narratives and by looking closely at heterogeneity in prosocial concerns.

Andreoni and Rao (2011) study a setting closely related to ours. In their experiment, Receivers and Dictators in a dictator game can communicate with each other. Andreoni and Rao (2011) find that giving increases whenever Receivers can say something. On the other hand, giving decreases when only Dictators can communicate. We investigate a setting that is related, but different, since active communication is absent in our experiment. In particular, our study speaks to situations in which a decision-maker can choose between alternatives that are more or less prosocial, but does not come into direct contact with those affected by her choice. She only listens to arguments of other decision-makers and has to take a decision.

Other work in this field has looked at how contextual factors, e.g., the role of frames (Brañas-Garza, 2007; Dreber et al., 2013) or social information (Krupka and Weber, 2009; Gino et al., 2009), influence prosocial behavior. We hold these channels constant and explicitly provide reasons, or narratives, for a certain action. This links our study to other papers investigating the effect of moral reminders or recommendations on behavior (see, e.g., Galbiati et al. (2008) on obligations and Croson and Marks (2001) on recommendations, both in the public-good game; or Mazar et al. (2008) in the context of lying). Most closely related to our paper is an experiment by Dal Bó and Dal Bó (2014), who look at the effect of moral suasion in the form of arguments in favor of the socially optimal contribution in a voluntary contribution game. In contrast to them, we look at a non-strategic setting where narratives can only affect preferences and cannot work as coordination devices. Moreover, our manipulation does not come directly from the experimenter, but is based on naturally occurring reasons subjects provide for their choices.[6] Last but not least, our type measure allows us to look at heterogeneous effects and to test what we call negative narratives more thoroughly..[7]

To achieve this goal, we make use of the SVO slider measure by Murphy et al. (2011) to measure social types. Variations of the SVO measure have been widely used in both psychology and economics to measure the heterogeneity in individual motives in social and moral dilemmas[8] (see Balliet et al., 2009, for a meta-study), e.g. in the public-good game (see e.g. Offerman et al., 1996). Grossman and Van Der Weele (2017) study a setting where people can remain ignorant about harmful consequences of their actions, and find that the SVO measure confirms the sorting predictions of their model. In line with previous studies, we are interested in how heterogeneous prosocial concerns interact with our treatment manipulations. We find that this is indeed an important dimension to look at, since different types display not only quantitatively, but also qualitatively different reactions.

[6] Dal Bó and Dal Bó (2014) (p. 30) explicitly encourage research on messages coming from other subjects and not from the experimenter.

[7] Dal Bó and Dal Bó (2014) find that messages explaining the game-theoretical prediction of zero contribution have no effect. However, baseline contributions are already quite low in their paper and there is hardly any room for a further decrease to take place.

[8] Other studies find that individuals scoring differently on the SVO measure exhibit different behavior also in other realms, such as intergroup conflict (Weisel et al., 2016), in vaccine-related behavior (Böhm et al., 2016), and in pay what you want settings (Krämer et al., 2017).

3 Experimental Design

3.1 Setup

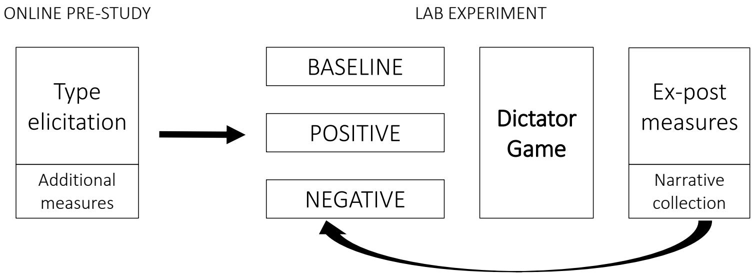


Figure 1: Experimental Design

Our experimental design consists of two main building blocks (see Figure 1), namely an online pre-study and a standard laboratory experiment. This experiment is subdivided in a modified *dictator game* and a questionnaire containing various ex-post measures. The online pre-study was conducted one week before the experiment.[9] The laboratory experiment was implemented in a between-subjects design with a Baseline and two treatment conditions (Positive and Negative), which varied the content of the narratives subjects saw. Below, we discuss the individual parts of the study in detail. Instructions for the laboratory experiment can be found in Appendix A.10.

Dictator game The central part of our design is constituted by a simple dictator game (Kahneman et al., 1986). Dictators chose how to divide 10 € between themselves and an anonymous recipient (in intervals of 1 €).

[9] Subjects received the link to the pre-study one week before the experiment and had three days to complete it.

All subjects in the experiment decided under role uncertainty,[10], i.e., each subject made her choice in the role of the dictator and roles were randomly assigned at the very end of the experiment. Instructions were neutrally framed, in that dictators were called "Participant A" and recipients "Participant B".

Crucially, we fixed the empirical expectations of subjects about the distribution of giving in dictator games. This ensures that subjects did not take narratives in our treatment conditions as signals about the empirical distribution of giving. To achieve this goal, subjects in all experimental conditions were presented with a graph showing the distribution of dictator game giving in similar experiments on the decision screen (see Figure 6 in Appendix A.1). This graph displays data from Engel (2011) restricted to studies in which 10 units of currency were used. The figure shows the typical bimodal distribution with most dictators giving either 5 € or 0 €. We told subjects that the graph displayed the distribution of choices other subjects had made in similar previous experiments.[11]

Treatments Participants were randomly allocated to one of three treatment conditions in a between-subjects design. In the BASELINE condition, subjects only saw the distribution of dictator game giving described above. In the two treatment conditions, they were additionally shown two comments that subjects in the BASELINE condition had used to explain their choices. These are our narratives. In the POSITIVE condition, subjects saw two comments in support of the equal split (giving 5 €), and in the NEGATIVE condition they saw two narratives justifying selfish behavior (giving 0 €). They were (truthfully) told that these were explanations other participants had given for their choices.[12] In the next paragraph, we explain how we collected and selected the narratives to devise our treatment conditions.

[10] Iriberri and Rey-Biel (2011) find that role uncertainty increases selfish choices. To the extent to which the increase is not excessive and does not interact with our treatment manipulations, this does not constitute a problem for our design.

[11] We used the following expression: "This figure shows the occurrence of choices of participants in similar experiments in percentages."

[12] We used the following expression: "Here are two explanations (*Begründungen*, in German), which other participants gave for their choice."

Narrative collection After subjects had gone through all stages of the experiment, but before their final roles for the payment were revealed, they were given the opportunity to explain the reasoning behind their choice in the dictator game.[13] We used the explanations from the Baseline condition to build the set of narratives that subjects saw in the other experimental conditions. Three independent raters, blind to the research question, evaluated the narratives along several dimensions. First, they were asked whether it was possible to understand, from the subject's comment, what he or she had chosen in the dictator game and, if so, which was the most likely choice (0,1,2, etc.). Raters also evaluated (on a 7-point Likert scale) how convincing they perceived the narrative to be.[14]

We then selected the most convincing narratives (taking average ratings) in support of giving 0 € and in support of giving 5 €. We excluded narratives which were particularly long or repetitive compared to the others. We selected four positive and four negative narratives. Each individual in each of the two treatment conditions saw two randomly selected narratives (at individual level). These steps were taken, on the one hand, to prevent our results from depending on a single item and, on the other, to increase the probability of subjects indeed being treated by at least one narrative. For more details on the procedure used to select the narratives, see Appendix A.2.

Type elicitation As mentioned above, the online pre-study was conducted one week prior to the laboratory experiment to avoid direct contamination effects. The purpose of our online pre-study was to measure the subjects' prosocial concerns to compare how different social types reacted to our treatment manipulations. Our main measure of a subject's social type is the SVO slider measure (Murphy et al., 2011). Subjects are confronted with 6 choices where they have to trade off their earnings with those of another subject under different budget constraints. From these

[13] The exact wording was the following. "You divided the money in the following way. You: €Ṗarticipant B: €. You can now explain ("*begründen*", in German) this decision for yourself." We asked subjects to stick to a maximum of two or three sentences and imposed a generous upper bound of 500 characters. This stage was unannounced.

[14] Additionally, raters evaluated the narratives with regard to their creativity, profoundness, and honesty. We do not use these measures in this study.

choices, the so-called SVO angle is constructed, which represents the relative weight subjects put on the income of others compared to their own income. Subjects with an SVO angle of 0° care only about their income, while those with an SVO angle of 45° weigh their income and that of the other subject equally. Types with an SVO angle below 25° are generally classified as selfish and those above as prosocials. Earnings in this task are determined by forming random pairs of subjects. One of the 6 choices is randomly selected and the choice of one of the two subjects in the pair is randomly implemented. For details on the measure, we refer to Murphy et al. (2011).

The SVO measure has been shown to be a stable and consistent predictor of behavior in different social dilemmas (see Balliet et al., 2009, for a meta-study). Moreover, high SVO types (**prosocials**) differ from low SVO types (**selfish**) in their decision-making process (e.g., Fiedler et al., 2013). This makes the SVO measure particularly suitable for capturing heterogeneity in reactions to our narrative manipulation.

We additionally elicit further psychological measures. We include the 11-item, Big5 questionnaire (Rammstedt and John, 2007), the Context Dependence and Independence questionnaire (Gollwitzer, 2006), a reduced form of the Moral Disengagement questionnaire (Bandura et al., 1996), and a modified version of the Moral Identity Scale (Aquino and Reed, 2002) (for more details on the measures we employ, see Appendix A.3). We use these measures (a) as controls in a robustness check of our treatment effects, and (b) to explore the role these psychological constructs play in explaining our treatment effect.

Ex-post measures Directly after the dictator game decision, subjects went through a series of stages meant to investigate potential mechanisms driving our treatment effects. We describe these questions in the order in which they were presented to participants.[15]

[15] We also asked subjects to state their personal norm, i.e., how much they thought would be appropriate to give. However, since the measure was elicited after subjects had made their choice, we cannot exclude that it was used in a self-serving manner to further justify their choice further. In fact, we find no clear variation between treatments and a high correlation with giving. For these reasons, we do not use this measure in our analysis.

1. General happiness and contentment.
2. Feelings with regard to dictator game choice (happiness, guilt, content, amusement, shame, pride, excitement).

Procedures The experiment was conducted at the DecisionLab of the Max Planck Institute for Research on Collective Goods in Bonn between May and June 2018. The online experiment was conducted using Qualtrics software, while the laboratory experiment was programmed in zTree (Fischbacher, 2007). Subjects were recruited via Orsee (Greiner, 2015). 282 participants (64% female, average age 24.8 years)[16] took part in the experiment. For the analysis, we exclude 2 subjects who had not taken part in the online pre-study. All subjects received a show-up fee of 5 €, plus their earnings from the the online pre-study (2 € participation fee plus between 0.50 € and 3 € for the SVO slider task) and their earnings from the dictator game. Overall, subjects received on average 14.48 €. The online pre-study lasted between 5 and 15 minutes, while the laboratory experiment took on average 40 minutes.

3.2 Hypotheses

We derive qualitative predictions for the effect of our treatment conditions from a simple model describing how giving is influenced by narratives.[17] Agents in our model choose how much money to give to someone else. The model is informed by Bénabou et al. (2018), where narratives are introduced as signals about the externality of a moral action (in our case: giving). A key component of this model is the belief about the externality, which we model as an environmental factor, and which can be interpreted as the "importance" or "goodness" of giving. We first describe the basic utility function of an agent; we then introduce the environmental factor; and finally, we discuss how narratives enter the model. Note that this model differs from more standard models, since heterogeneity will follow solely from heterogeneous beliefs about the externality or goodness of giving and not from a given preferences parameter about giving. We will introduce

[16] For 74 subjects, this information was not recorded.

[17] We thank Arno Apffelstaedt for the initial modeling idea.

narratives as a signal about the environmental factor and consider a standard Bayesian framework for the updating process. These assumptions lead to a tractable and straightforward model that allows us to generate two main predictions for our setting.

The utility function of an agent takes the following form:

$$U_i(g, e) = v(g, e) - c(g), \tag{1}$$

where g is the amount she decides to give, and e is an environmental factor which we describe below. $v(g, e)$ captures the overall valuation of giving and $c(g)$ the costs of giving. We set $e \in \{0, 1\}$ and assume $c(g)$ to be linear in g. While $v(g, e)$ can take many functional forms, we assume concavity in g ($\frac{\partial v(g,e)}{\partial g} > 0$, $\frac{\partial^2 v(g,e)}{\partial g^2} < 0$). This assumption ensures an internal solution with an optimal amount of giving $g^*(e)$.

The environmental factor The environmental factor or externality e is binary and measures the presence of an externality or the appropriateness of giving in the situation at hand. If $(e = 1)$, there is a positive externality or it is appropriate to give, while if $(e = 0)$, there is no such externality or giving is not required in this situation. The idea is that as an agent, say, a donor, reaps higher benefits from giving when the recipient is indeed deserving or needy. We assume that the marginal utility of giving is increasing in the environmental factor e ($\frac{\partial v(g,e)/\partial g}{\partial e} > 0$). Following this assumption, a higher e leads to higher amounts of giving. Note that $v(g, e)$ can take on many different forms. In a setting like the standard dictator game, the strong focal point at the equal split, could be captured with a normative component n in the utility function. As an example, setting $v(g, e) = -\gamma(e)(n - g)^2$, e would capture the appropriateness to follow the norm (assuming $\frac{\partial \gamma}{\partial e} > 0$). Independently of the specific choice of v, our predictions hold.

Agents in our model do not know the value of e with certainty. Rather, agents hold initial, prior beliefs about e with $\hat{e} = P(e = 1)$. These beliefs can be understood as the extent to which an individual thinks it is appropriate to give, or her beliefs about the value of the positive externality arising from giving (Bénabou et al., 2018). Agents who perceive the externality as low might believe that the recipient is not deserving or that it is not appropriate or necessary to give in that specific situation. This makes these agents less prone to give compared to agents perceiving the externality as high or believing that it is appropriate to give in that situation. An individual's belief may be deeply grounded in her, it may have formed through experience, or, alternatively, she may self-servingly hold a belief that allows her to act selfishly. From the above assumptions, it follows naturally that higher beliefs about $e = 1$ lead to higher amounts of giving.

Heterogeneity Decision-makers in our model differ solely in their beliefs about e, which we bound to $\hat{e} \in (0, 1)$. That is, all decision-makers in our model would act in the same way, i.e., choose to give the same amount of giving, if they held the same belief.

Modelling choices through beliefs about e grants flexibility to our model in that beliefs are not fixed, i.e., types are malleable. Decision-makers with a low belief about $e = 1$ (small $\hat{e}$), i.e., selfish types, can be persuaded to update their beliefs upwards and give more. Prosocial types, i.e., decision-makers with high $\hat{e}$, on the other hand, can be persuaded that it is not necessary to give. This allows an application of our model to many different contexts where one would expect environmental factors to influence choices, e.g., framing. Most importantly for this paper, we can introduce narratives as a signal about the externality with potentially different effects for different subjects.

It is important to stress the link between our model and more classical preference models. To illustrate this, take the SVO measure that we use in the experiment to classify subjects into types. SVO is a measure of prosociality that correlates with choices in moral or cooperation dilemmas. A higher SVO angle represents a higher weighting of the payoffs of others in a standard social preference model. In a neutral, unperturbed setting like the dictator game in our Baseline, we would assume, following our model, that beliefs translate directly to choices. Thus, the SVO measure should be a good proxy of beliefs about the externality as described by our model. That is, one would expect selfish types to have a low belief about e. Prosocial types might interpret the same situation quite differently and have a higher belief about e, i.e., they might think that it is important or necessary to give or else that there is a positive externality.

Narratives We model narratives as signals about e updating the prior belief of a decision-maker, as in Bénabou et al. (2018). A positive narrative will signal that $e = 1$, i.e., it is an argument or justification for why it is appropriate to give in the situation or for there being a positive externality. A negative narrative, conversely, will signal that $e = 0$. For simplicity, we take decision-makers to be standard Bayesian updaters.[18] Further, we assume that narratives are at least somewhat believable, which in the model translates to the signal being correct more often than not.[19] Given this signal structure, negative narratives lead to a downward shift in beliefs and positive narratives to an upward shift (see Appendix A.4 for a detailed example of the updating process). That is, independent of the prior belief, the posterior belief is decreasing when receiving a negative narrative and increasing when receiving a positive narrative for the full range of beliefs. Since, as stated above, higher beliefs about e translate into higher amounts of giving, our first hypothesis follows directly.

Hypothesis 1 *Positive narratives increase giving, while negative narratives decrease giving.*

While in our model the overall effect of narratives is independent from the prior belief, i.e., the social type of a subject, our setup predicts a different strength of the effect for different social types. In particular, extreme types (those with priors $\hat{e}$ close to 0 or close to 1) will not update strongly when receiving a signal close to their prior belief, whereas they will update strongly when receiving a contradicting signal. That is, selfish subjects (low $\hat{e}$) should be strongly influenced by positive narratives and

[18] Other forms of updating are of course conceivable, but would introduce further degrees of freedom to the model and are not a priori clearly defensible. Moreover, as long as an alternative updating model leads to updating in the same direction for all priors and leads to different posteriors for different priors, the main intuitions of the model will hold.

[19] Note that it is sufficient if the decision-maker *perceives* the narrative as believable. In the experiment, we take care of this by selecting only the narratives perceived as most convincing by independent raters.

prosocial subjects (high $\hat{e}$) should be more strongly influenced by negative narratives (see the example in Figure 8 in Appendix A.4).

Hypothesis 2 *Positive narratives should have the strongest positive effect on more selfish subjects, while negative narratives should have the strongest negative effect on more prosocial subjects.*

Importantly, the goal of this model is not to make point predictions for specific types. Clearly this would require an assumption on the precision of the signal and on the range of prior beliefs, as well as on how exactly the SVO measure maps into the prior beliefs. Rather, this model provides a benchmark to judge the effects of our treatments.

4 Results

Our dataset consists of 280 independent observations spread over three treatment conditions. In the first part of this section, we analyze the evidence regarding the hypotheses derived from the theoretical model above. We then provide some additional results that will help guide our discussion.

4.1 Main results

In the Baseline condition, subjects give on average 2.76 €. According to Hypothesis 1, we should observe an increase in average giving in the Positive condition and a decrease in the Negative condition. Figure 2 provides a visual representation of the aggregate results. In the Positive condition, average giving increases to 3.23 €. This constitutes a 17% increase, in line with our first hypothesis. The difference, however, is only marginally significant (rank-sum test, $p = .0932$). Average giving in the Negative condition is virtually identical, at 2.78 €, to the level of giving in the Baseline condition (rank-sum test, $p = .9076$).

However, these aggregate results on giving provide an incomplete picture of the data. As stated in Hypothesis 2, social types closer to the extremes should respond more strongly to our treatment manipulations. Although, according to our theoretical model, the effect should go in the same direction for all types.

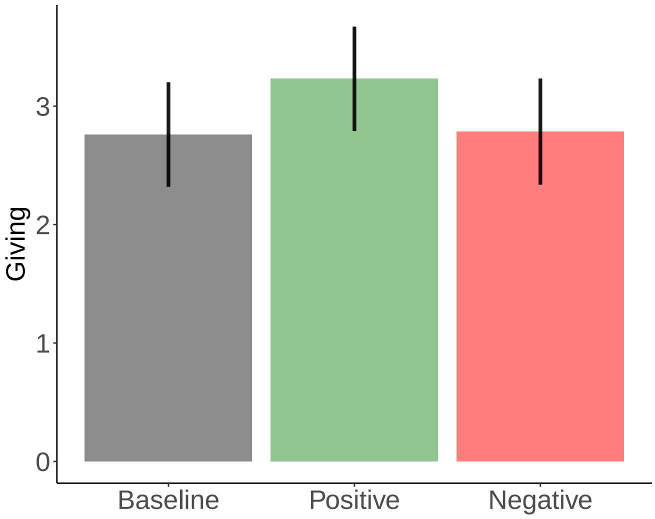


Figure 2: Average giving across treatments with 95%-confidence intervals.

Figure 3 displays the relationship between an individual's giving choices and her social type. Giving is, as is typical in dictator games, bounded above at 5 € with only two subjects giving 6 € and many giving nothing at all. We use LOESS fitted lines to provide a better visualization of the data. The black solid line depicts the relationship between type and giving in Baseline; the green dotted line represents our Positive condition and the red dashed line our Negative condition. We observe the expected positive correlation between our measure of type and giving in the Baseline condition. The steepness of the fitted line in the middle of the graph indicates that, in line with previous studies (see Engel, 2011), giving follows a bimodal distribution, with most subjects deciding to give either half of their endowment or nothing.

To test how different types react to different narratives, we run a Tobit regression with the amount of giving as the dependent variable and treatment dummies, type, and interaction terms between type and treatment dummies as explanatory variables (see Table 1).

We first look at column (1), where we introduce our treatment conditions as dummies and control for the social type of a subject. This shows a strong positive effect of the Positive condition on giving, confirming part of Hypothesis (1). The overall effect of the Negative condition is also positive, but small and clearly not significant. Note that, as expected, the

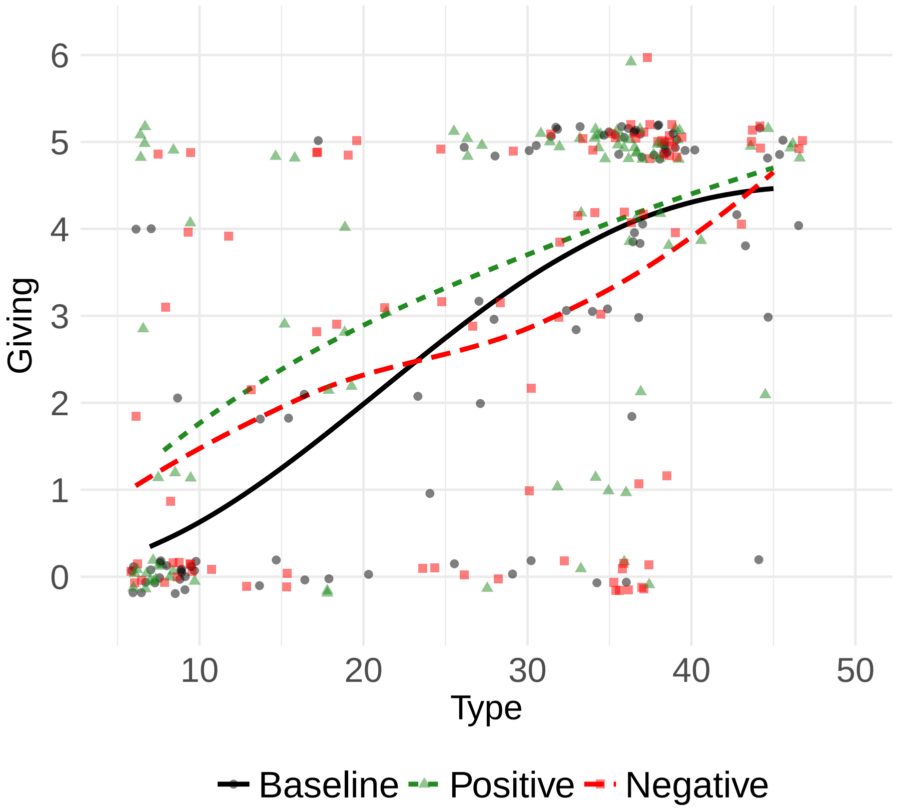


Figure 3: Giving on SVO. LOESS fitted lines.

Note: Data points are jittered. For the ease of visualization, we removed social types below 5° and above 50°, which are rare and not balanced across treatments.

type measure is a clear predictor of giving: the higher the SVO angle of a subject is, the more she gives.

Adding an interaction between social type and treatment condition in column (2), we find a positive effect of both conditions on giving for more selfish types. The interaction terms reveal that both the effect of the Positive and that of the Negative condition decrease as the social type of a subject increases, i.e., for more prosocial types.

We turn to Figure 4 for a clearer interpretation of the interaction terms.[20] Here we plot the predicted marginal effects of our treatment conditions on giving compared to the Baseline. This enables us to get a qualitative test of Hypothesis 2.

[20] For the distribution of types split by condition, see Figure 9 in Appendix A.5.

dv: giving	(1)	(2)
POSITIVE	0.752**	2.852***
	(2.09)	(3.21)
NEGATIVE	0.125	2.698***
	(0.35)	(3.02)
Type	0.133***	0.189***
	(11.41)	(8.74)
POSITIVE x type		-0.0732**
		(-2.58)
NEGATIVE x type		-0.0900***
		(-3.16)
Constant	-1.382***	-3.015***
	(-3.23)	(-4.33)
Observations	280	280
Pseudo R^2	0.108	0.118

t statistics in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

Table 1: Tobit regressions.

Note: Coefficients of Tobit regression with lower censoring at 0. The type measure corresponds to the SVO angle, Positive and Negative conditions are introduced as dummies. We also include interaction terms between conditions and types.

We start with the Positive condition (green dotted line), where we find a pattern in line with our hypothesis. We notice a strong positive effect for more selfish types, which fades out for more prosocial types. Evaluating this effect for the modal selfish type (60 subjects with an SVO angle of 7.82°), this translates in a positive and significant difference of 2.28 € ($p = .001$) in giving. Prosocial types, on the other hand, display no significant increase.

Result 1 (Positive Narratives) *Positive narratives increase giving compared to the* Baseline *condition, especially for selfish subjects.*

The Negative condition (red dashed line), instead, generates a qualitatively different pattern. More selfish types increase their giving compared to the Baseline. The difference of 2 € ($p = .004$) for the modal selfish type is positive and significant. Note that this increase is statistically indistinguishable from that of the Positive condition. This effect is clearly not

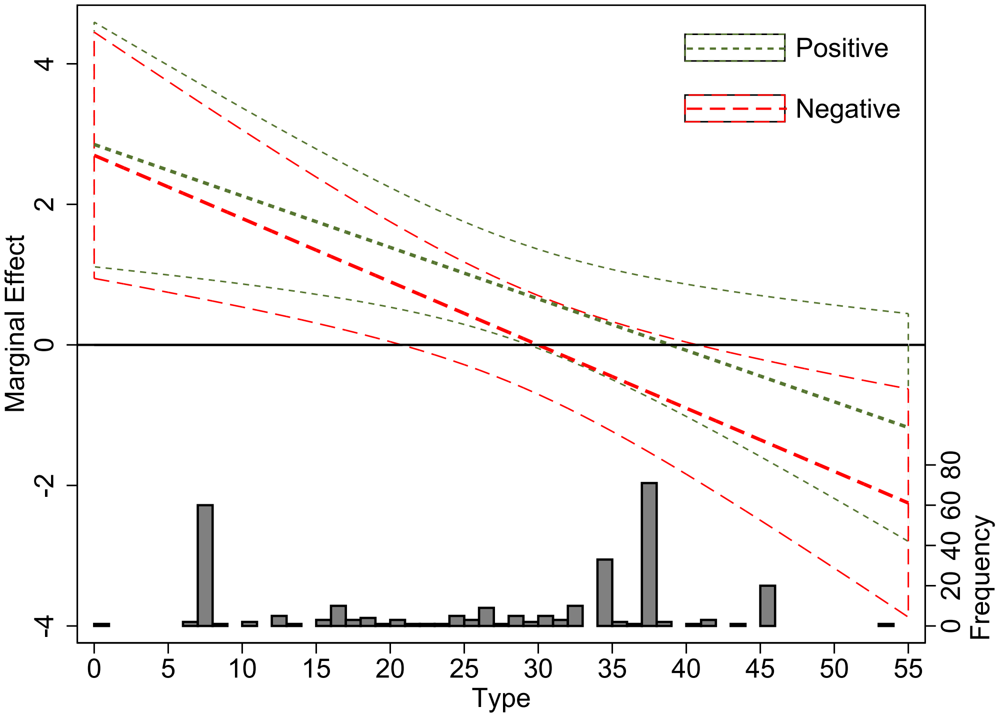


Figure 4: Marginal effects on types, 95% confidence intervals.

Note: In the lower part of the graph, we plot the pooled distribution of types over all conditions. For visualization, types below 0° (3 subjects) are removed.

in line with the pattern predicted in Hypothesis 2. More prosocial types, on the other hand, give less than in the BASELINE. The modal prosocial type (61 subjects with an SVO angle of 37.48°) decreases giving by 0.67 € ($p = .121$), which is not statistically significant. However, for more prosocial types (21 subjects with an SVO angle above 44°), the effect becomes negative and significant.

Result 2 (Negative Narratives) *Negative narratives have a differential effect: they decrease giving for prosocial types and increase giving for selfish types compared to the* BASELINE.

We run further regressions to check the robustness of our results (see Appendix A.6). First, we compare the results from the Tobit regressions with a standard OLS regression. We then include the additional psychological measures collected in the online pre-study as controls to our Tobit model. We also run the same specification of Model (2) in Table 1 using both upper and lower censoring. Finally, we include a quadratic interaction term between our treatment conditions and the social type to capture potential nonlinearities. Our results are robust to this additional analysis.[21]

4.2 Additional results

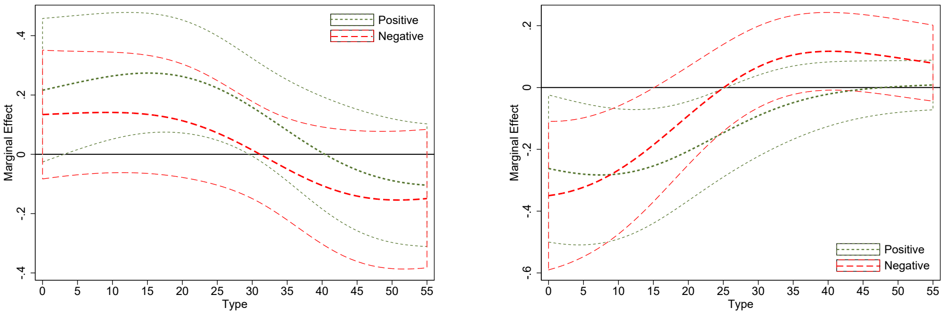


Figure 5: Marginal effects, probit

Note: Regressions with dummy of giving 5 (left) and giving 0 (right). Including interaction terms. Outer lines show 95 % confidence intervals. For visualization, types below 0° (3 subjects) are removed.

We round up this section by looking at additional results that go beyond the predictions of our theoretical model. This will help us to give a clearer and more encompassing interpretation of our data in the discussion. A natural question is whether subjects who changed their behavior did so by going all the way to the action prescribed by the narratives or whether they only partially changed their decision. In other words, did the Positive (Negative) condition lead subjects to give 5 (0) more frequently than in the Baseline?

To answer this question, we run two Probit models on the probability of giving either 5 or 0. The graphs in Figure 5 show the predicted marginal effects on different social types for the same specification we used in the

[21] We also perform our analysis using the Moral Identity Scale and Moral Disengagement as alternatives to the SVO angle. Both have a strong and stable relationship with giving, but turn out to be irrelevant in explaining our treatment effect. Moreover, Context Dependence or Independence do not mediate our treatment effects. This gives us further assurance in using the SVO as our type measure for the main analysis (see Appendix A.7).

Tobit model above (see Table 4 in Appendix A.8 for the full regression). There are three main observations to be made. First, the probability of selfish types giving an amount equal to 5 in the Positive condition increases for nearly all selfish types.[22] This translates into a 26% ($p = .022$) increase in the probability of giving 5 for our modal selfish type (SVO angle of 7.82°) in the Positive condition. The same is not true for the Negative condition, where the increase in the probability of giving 5 is smaller and statistically insignificant. The difference for the modal selfish type is just 14% and not significant ($p = .178$). Second, both in the Positive and the Negative condition the probability of selfish types giving 0 decreases substantially. This effect is observed for a larger type range in the Positive condition. The decrease in the probability of giving 0 for the modal selfish type (SVO angle of 7.82°) corresponds to 28% ($p = .012$) and 30% ($p = .007$) in the Positive and Negative condition, respectively. Third, we find that, although more prosocial types give less in the Negative condition, this does not substantially increase the probability of giving 0 for them. The increase in probability for the modal prosocial type (SVO angle of 37.48°) is moderate (11%) and only marginally significant ($p = .077$).

Result 3 *The* Positive *condition increases the probability of following the narrative (giving 5) for selfish types and both treatment conditions decrease the probability that selfish types give 0. The latter effect applies to a wider range of types for the* Positive *condition. The* Negative *condition does not lead to a substantial increase in the probability of giving 0 for prosocial types.*

We finally look at the effect of our treatment conditions on the ex-post measures of subjects' feelings (Table 5 in Appendix A.9 shows our regression analysis). We find no treatment effects on general happiness or contentment. Feelings of guilt and shame with regard to the choices made by subjects have, as one could expect, a strong and stable relation with the amount of giving: giving less increases these reported feelings. However, our treatment conditions do not increase or reduce guilt or shame about choices. Nevertheless, we cannot rule out that the absence of treatment

[22] The effect is particularly strong for the range of selfish types who are more frequent in our sample (those above an SVO angle of 5° and below one of 25°).

effects is due to an anticipation of these feelings. The presence of narratives could lead subjects to anticipate these feelings and to adapt their giving to avoid them, which could result in similar levels of uilt and shame across treatments.

Result 4 *The effect of our treatment conditions does not seem to operate by directly changing subjects' feelings towards their choice.*

5 Discussion and Conclusion

Our results provide new insights into how narratives in favor of prosocial or selfish actions influence the behavior of different social types. Subjects in our experiment see either positive or negative narratives upon taking a distributional choice in a dictator game. We compare these two treatment conditions with a baseline in which no such narratives are provided. Empirical beliefs about the distribution of choices are fixed across all experimental conditions. We derive two general hypotheses from a basic model of how narratives influence behavior via beliefs about the externality or appropriateness of an action for different social types.

Subjects in the Positive condition give more than subjects in the Baseline condition. This increase is predominantly driven by selfish types (Result 1). On the other hand, narratives in the Negative condition have a striking differential effect (Result 2). Prosocial types in the Negative condition give less than in the Baseline. However, this effect is reversed for selfish types, who give more in the Negative condition compared to the Baseline, matching the giving level of their peers in the Positive condition. These results are only partly in line with the hypotheses derived from our model. In particular, our model allows social types to react to different extents to narratives, but assumes that the effect of the narrative goes in the same direction for all social types.

The differential effect of narratives resonates well with other research that provides evidence that different social types process information differently (Fiedler et al., 2013), have a different representation of moral dilemmas (Van Lange et al., 1990; Liebrand et al., 1986), and that, more in general, prosocial and selfish decisions are qualitatively different (Rand et al., 2012). Taken together, this suggests that the behavior of prosocial and selfish types in our experiment might have been driven by different motivations.

We suggest two potential candidates: an enhanced saliency effect of narratives and a social comparison argument. According to the first explanation, as we have just pointed out, the more selfish subjects might not consider the moral consequences of their actions in their "ordinary" decision process. This could be due to them genuinely not being aware of these consequences or self-servingly suppressing them. In this case, the mere presence of a narrative, regardless of its content, could make the moral nature of the situation more salient. This would lead selfish subjects to give more. This conjecture is in line with a study by Krupka and Weber (2009), who find that normative information enhances prosocial behavior, even in cases where one does not expect or does not observe a lot of norm-compliant behavior. Similarly, Gino et al. (2009) find that increasing the saliency of an opportunity to cheat decreases unethical behavior.

The second explanation relies on the subjects comparing themselves with the narrator. If subjects care about their self-image, i.e., how they view themselves, the content of the narrative could serve as a social reference point. In particular, narratives in the Negative condition would represent a very low reference point. Giving at least something after facing a negative narrative provides a low-cost opportunity for a selfish type to distinguish herself from the narrator. The reverse would hold for prosocial types. Since their reference point is very low by comparison, they can decrease their giving slightly, while withholding a positive image of themselves. This account is coherent with psychological theories underlying the importance of social comparison for people's self-perception (Festinger, 1954) and also with empirical evidence that people indeed adapt their behavior accordingly (Frey and Meier, 2004).

Our study was not designed to distinguish between these two mechanisms and we can only offer suggestive evidence in favor of one or the other. The results we obtain from a Probit regression with either the equal split or the selfish action as dependent variables (Result 3) do not refute either of the two explanations. Subjects in the Negative condition are not pushed to the extremes: the probability that prosocial (selfish) types give nothing (split equally) does not substantially increase. These patterns are compatible with the phenomenon of partial lying (Fischbacher and Föllmi-Heusi, 2013) or ethical manoeuvring (Mazar et al., 2008; Shalvi et al., 2011b), which is consistently found in the literature on lying and cheating. Subjects in those experiments do not lie to the full extent, in order to avoid being unequivocally identified as liars or cheaters. Likewise, prosocial subjects who buy into a negative narrative do not go all the way to giving nothing at all. Conversely, while selfish subjects are less likely to give nothing at all, they do not completely switch to the equal split. The picture is different in the Positive condition. Here, positive narratives seem to induce a different type of moral trade-off. Selfish subjects are drawn more strongly to the action prescribed by the narrative, as implied by the increased probability of equal split divisions. Hence, narratives in favor of the moral action, it appears, imply a binary response (comply or not comply) instead of a more gradual trade-off, as in the case of negative narratives.

Our work has relevant implications for understanding the determinants of prosocial and moral behavior. Much work has focused on the role that structural or contextual factors, e.g., incentives or empirical information, play in moral trade-offs. However, narratives permeate every aspect of human behavior and are fundamental tools that guide people's actions. This study shows that these factors can have a sizeable impact on economic behavior. Moreover, the differential effect we find highlights that it is crucial to take into account potential heterogeneity when evaluating aggregate outcomes. This is true both for experimental studies and for policy interventions. Acknowledging the importance of narratives for economic behavior can help to enrich the understanding of various phenomena, offering important prospects for policy design and intervention.

There are new questions arising from our work and more insights are needed to gain a more comprehensive understanding of these phenomena. Further studies should try to obtain insights on the different mechanisms that seem to be at play for selfish and prosocial types. More work is also needed to determine the effectiveness of different narratives in influencing behavior. Further questions are how enduring the effect of a certain narrative is and whether it might have spillovers in other contexts. From a theoretical point of view, a model with the aim of accounting the evidence presented here would need to capture the differential effect of narratives on different social types. We hope that our work can contribute to inspire such endeavors.

References

Akerlof, George A and Dennis J Snower, "Bread and bullets," *Journal of Economic Behavior & Organization*, 2016, *126*, 58–71.

_ **and William T Dickens**, "The economic consequences of cognitive dissonance," *The American Economic Review*, 1982, *72* (3), 307–319.

Andreoni, James and Justin M Rao, "The power of asking: How communication affects selfishness, empathy, and altruism," *Journal of Public Economics*, 2011, *95* (7-8), 513–520.

Aquino, Karl and I.I. Reed, "The self-importance of moral identity.," *Journal of Personality and Social Psychology*, 2002, *83* (6), 1423.

Balliet, Daniel, Craig Parks, and Jeff Joireman, "Social value orientation and cooperation in social dilemmas: A meta-analysis," *Group Processes & Intergroup Relations*, 2009, *12* (4), 533–547.

Bandura, Albert, Claudio Barbaranelli, Gian Vittorio Caprara, and Concetta Pastorelli, "Mechanisms of moral disengagement in the exercise of moral agency.," *Journal of Personality and Social Psychology*, 1996, *71* (2), 364.

Bénabou, Roland and Jean Tirole, "Mindful economics: The production, consumption, and value of beliefs," *Journal of Economic Perspectives*, 2016, *30* (3), 141–64.

_ **, Armin Falk, and Jean Tirole**, "Narratives, Imperatives and Moral Reasoning," 2018.

Bicchieri, Cristina and Hugo Mercier, "Self-serving biases and public justifications in trust games," *Synthese*, 2013, *190* (5), 909–922.

Bó, Ernesto Dal and Pedro Dal Bó, ""Do the right thing:" The effects of moral suasion on cooperation," *Journal of Public Economics*, 2014, *117*, 28–38.

Böhm, Robert, Cornelia Betsch, and Lars Korn, "Selfish-rational non-vaccination: experimental evidence from an interactive vaccination game," *Journal of Economic Behavior & Organization*, 2016, *131*, 183–195.

Brañas-Garza, Pablo, "Promoting helping behavior with framing in dictator games," *Journal of Economic Psychology*, 2007, *28* (4), 477–486.

Bruner, Jerome, "The narrative construction of reality," *Critical Inquiry*, 1991, *18* (1), 1–21.

Carlson, Ryan W, Michel Marechal, Bastiaan Oud, Ernst Fehr, and Molly Crockett, "Motivated misremembering: Selfish decisions are more generous in hindsight," 2018.

Chance, Zoë, Michael I Norton, Francesca Gino, and Dan Ariely, "Temporal view of the costs and benefits of self-deception," *Proceedings of the National Academy of Sciences*, 2011, *108* (Supplement 3), 15655–15659.

Croson, Rachel and Melanie Marks, "The effect of recommended contributions in the voluntary provision of public goods," *Economic Inquiry*, 2001, *39* (2), 238–249.

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

Ditto, Peter H, David A Pizarro, and David Tannenbaum, "Motivated moral reasoning," *Psychology of Learning and Motivation*, 2009, *50*, 307–338.

Dreber, Anna, Tore Ellingsen, Magnus Johannesson, and David G Rand, "Do people care about social context? Framing effects in dictator games," *Experimental Economics*, 2013, *16* (3), 349–371.

Engel, Christoph, "Dictator games: A meta study," *Experimental Economics*, 2011, *14* (4), 583–610.

Epley, Nicholas and Thomas Gilovich, "The mechanics of motivated reasoning," *Journal of Economic Perspectives*, 2016, *30* (3), 133–40.

Feiler, Lauren, "Testing models of information avoidance with binary choice dictator games," *Journal of Economic Psychology*, 2014, *45*, 253–267.

Festinger, Leon, "A theory of social comparison processes," *Human Relations*, 1954, *7* (2), 117–140.

_ , *A theory of cognitive dissonance*, Vol. 2, Stanford university press, 1962.

Fiedler, Susann, Andreas Glöckner, Andreas Nicklisch, and Stephan Dickert, "Social value orientation and information search in social dilemmas: An eye-tracking analysis," *Organizational Behavior and Human Decision Processes*, 2013, *120* (2), 272–284.

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

_ **and Franziska Föllmi-Heusi**, "Lies in disguise—an experimental study on cheating," *Journal of the European Economic Association*, 2013, *11* (3), 525–547.

Foerster, Manuel and Joel J van der Weele, "Denial and Alarmism in Collective Action Problems," 2018.

Frey, Bruno S and Stephan Meier, "Social comparisons and prosocial behavior: Testing" conditional cooperation" in a field experiment," *American Economic Review*, 2004, *94* (5), 1717–1722.

Galbiati, Roberto, Pietro Vertova et al., "Obligations and cooperative behaviour in public good games," *Games and Economic Behavior*, 2008, *64* (1), 146–170.

Gino, Francesca, Michael I Norton, and Roberto A Weber, "Motivated Bayesians: Feeling moral while acting egoistically," *Journal of Economic Perspectives*, 2016, *30* (3), 189–212.

_ **, Shahar Ayal, and Dan Ariely**, "Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel," *Psychological Science*, 2009, *20* (3), 393–398.

_ , _ , and _ , “Self-serving altruism? The lure of unethical actions that benefit others,” *Journal of Economic Behavior & Organization*, 2013, *93*, 285–292.

Golman, Russell, George Loewenstein, Karl Ove Moene, and Luca Zarri, “The preference for belief consonance,” *Journal of Economic Perspectives*, 2016, *30* (3), 165–88.

Greiner, Ben, “Subject pool recruitment procedures: organizing experiments with ORSEE,” *Journal of the Economic Science Association*, 2015, *1* (1), 114–125.

Grossman, Zachary and Joel J Van Der Weele, “Self-image and willful ignorance in social decisions,” *Journal of the European Economic Association*, 2017, *15* (1), 173–217.

Haisley, Emily C and Roberto A Weber, “Self-serving interpretations of ambiguity in other-regarding behavior,” *Games and Economic Behavior*, 2010, *68* (2), 614–625.

Hamman, John R, George Loewenstein, and Roberto A Weber, “Self-interest through delegation: An additional rationale for the principal-agent relationship,” *American Economic Review*, 2010, *100* (4), 1826–1846.

Iriberri, Nagore and Pedro Rey-Biel, “The role of role uncertainty in modified dictator games,” *Experimental Economics*, 2011, *14* (2), 160–180.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler, “Fairness and the assumptions of economics,” *Journal of Business*, 1986, pp. S285–S300.

Karlsson, Niklas, George Loewenstein, Jane McCafferty et al., “The economics of meaning,” *Nordic Journal of Political Economy*, 2004, *30* (1), 61–75.

Konow, James, “Fair shares: Accountability and cognitive dissonance in allocation decisions,” *American Economic Review*, 2000, *90* (4), 1072–1091.

Krämer, Florentin, Klaus M Schmidt, Martin Spann, and Lucas Stich, "Delegating pricing power to customers: Pay what you want or name your own price?," *Journal of Economic Behavior & Organization*, 2017, *136*, 125–140.

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

Krupka, Erin L and Roberto A Weber, "Identifying social norms using coordination games: Why does dictator game sharing vary?," *Journal of the European Economic Association*, 2013, *11* (3), 495–524.

Lange, Paul AM Van, Wim BG Liebrand, and D Michael Kuhlman, "Causal attribution of choice behavior in three N-person prisoner's dilemmas," *Journal of Experimental Social Psychology*, 1990, *26* (1), 34–48.

Larson, Tara and C Monica Capra, "Exploiting moral wiggle room: Illusory preference for fairness? A comment," *Judgment and Decision Making*, 2009, *4* (6), 467.

Liebrand, Wim BG, Ronald WTL Jansen, Victor M Rijken, and Cor JM Suhre, "Might over morality: Social values and the perception of other players in experimental games," *Journal of Experimental Social Psychology*, 1986, *22* (3), 203–215.

M., Schmidthals K. Pöhlmann C. Gollwitzer, "Relationalitäts-Kontextabhängigkeits-Skala (RKS): Entwicklung und erste Ansätze zur Validierung. (Berichte aus der Arbeitsgruppe "Verantwortung, Gerechtigkeit, Moral" Nr. 161)," *Trier: Universität Trier*, 2006.

Matthey, Astrid and Tobias Regner, "Do I really want to know? A cognitive dissonance-based explanation of other-regarding behavior," *Games*, 2011, *2* (1), 114–135.

Mazar, Nina, On Amir, and Dan Ariely, "The dishonesty of honest people: A theory of self-concept maintenance," *Journal of Marketing Research*, 2008, *45* (6), 633–644.

McAdams, Dan P, *Power, intimacy, and the life story: Personological inquiries into identity*, Guilford Press, 1988.

Murphy, Ryan, Kurt Ackermann, and Michel Handgraaf, “Measuring social value orientation,” *Judgment and Decision Making*, 2011, *6* (8), 771–781.

Offerman, Theo, Joep Sonnemans, and Arthur Schram, “Value orientations, expectations and voluntary contributions in public goods,” *The Economic Journal*, 1996, pp. 817–845.

Rammstedt, Beatrice and Oliver P John, “Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German,” *Journal of Research in Personality*, 2007, *41* (1), 203–212.

Rand, David G, Joshua D Greene, and Martin A Nowak, “Spontaneous giving and calculated greed,” *Nature*, 2012, *489* (7416), 427.

Rodriguez-Lara, Ismael and Luis Moreno-Garrido, “Self-interest and fairness: self-serving choices of justice principles,” *Experimental Economics*, 2012, *15* (1), 158–175.

Saucet, Charlotte and Marie Claire Villeval, “Motivated Memory in Dictator Games,” 2018.

Shalvi, Shaul, Francesca Gino, Rachel Barkan, and Shahar Ayal, “Self-serving justifications: Doing wrong and feeling moral,” *Current Directions in Psychological Science*, 2015, *24* (2), 125–130.

_ , Jason Dana, Michel JJ Handgraaf, and Carsten KW De Dreu, “Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior,” *Organizational Behavior and Human Decision Processes*, 2011, *115* (2), 181–190.

_ , Michel JJ Handgraaf, and Carsten KW De Dreu, “Ethical manoeuvring: Why people avoid both major and minor lies,” *British Journal of Management*, 2011, *22*, S16–S27.

Shiller, Robert J, "Narrative economics," *American Economic Review*, 2017, *107* (4), 967–1004.

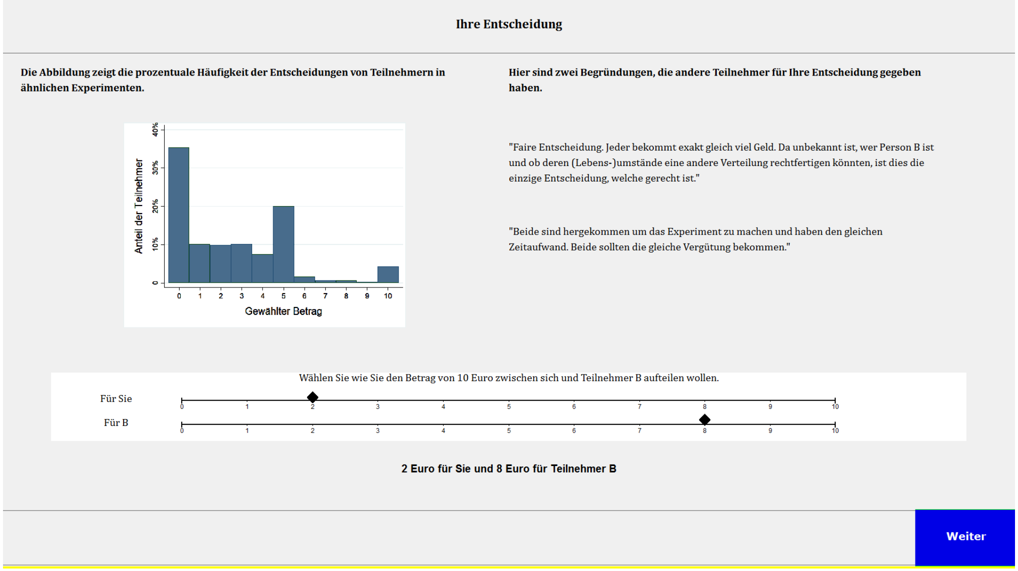
van der Weele, Joël J, Julija Kulisa, Michael Kosfeld, and Guido Friebel, "Resisting moral wiggle room: how robust is reciprocal behavior?," *American Economic Journal: Microeconomics*, 2014, *6* (3), 256–64.

Weisel, Ori et al., "Social motives in intergroup conflict: Group identity and perceived target of threat," *European Economic Review*, 2016, *90*, 122–133.

Wiltermuth, Scott S, "Cheating more when the spoils are split," *Organizational Behavior and Human Decision Processes*, 2011, *115* (2), 157–168.

A Appendix

A.1 Decision Screen



Ihre Entscheidung

Die Abbildung zeigt die prozentuale Häufigkeit der Entscheidungen von Teilnehmern in ähnlichen Experimenten.

Hier sind zwei Begründungen, die andere Teilnehmer für Ihre Entscheidung gegeben haben.

"Faire Entscheidung. Jeder bekommt exakt gleich viel Geld. Da unbekannt ist, wer Person B ist und ob deren (Lebens-)umstände eine andere Verteilung rechtfertigen könnten, ist dies die einzige Entscheidung, welche gerecht ist."

"Beide sind hergekommen um das Experiment zu machen und haben den gleichen Zeitaufwand. Beide sollten die gleiche Vergütung bekommen."

Wählen Sie wie Sie den Betrag von 10 Euro zwischen sich und Teilnehmer B aufteilen wollen.

Für Sie

Für B

2 Euro für Sie und 8 Euro für Teilnehmer B

Figure 6: Dictator game decision screen

Note: The decision screen shows the empirical distribution of choices on the left. On the right side the two narratives are listed. At the bottom, the subject is prompted to make his choice with the use of two sliders marking the payment for the subject and the recipient (the amounts on the sliders always add up to 10 €).

A.2 Narrative Selection

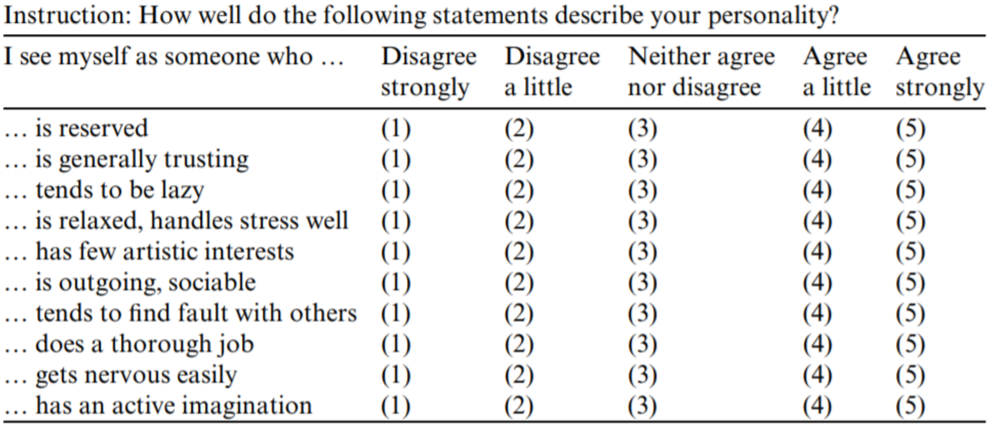
The following table shows positive and negative narratives along with their average convincingness rating. Numbers 1-4 were selected for the Positive condition and 5-8 for the Negative condition. Narratives were selected from all narratives of the first 3 sessions of the Baseline condition, since the 4th session was run later to balance the number of participants in all conditions. We took only narratives for which all raters were sure that they came from subjects giving 5 € or 0 €, respectively. We singled out the ones with the highest convincingness rating and removed repetitive ones and some that were considerably longer or shorter than the others.

Number	Positive Narratives	Convincingness
1	Both came here to participate in the experiment and spent the same amount of time here. Both should get the same payment.	6
2	An equal distribution of the money is only logical: Assuming everyone agrees on that, everyone will go home with 10 €. Everything else would be a mixture of greed and speculation.	6
3	Fair choice. Everyone gets exactly the same amount of money. Since it is unknown who Person B is and whether her life circumstances would justify another distribution, this is the only just decision.	6
4	I think that both participants should get the same amount of money. If it is unknown in advance whether you are A or B it is just smart to give 5 € to both.	6.3
	Negative Narratives	
5	Since the experiment is anonymous, I expect that everyone is looking for her own advantage. I don't know any of the other players and since the decision happens randomly anyway, I do not care about giving someone else money.	6
6	This way I get the highest payoff in case I am participant A. In case I am participant B, I have no influence on my payoff because of the assignment to participant B.	5.6
7	Because I would like to have the money and saw in the statistic that others also decided this way. This made me have less scruples for allocating all the money to myself.	5.3
8	I allocated 10 € to myself, since this way I get the most money on average. As it is unclear how much I would get as participant B, I wanted to achieve the maximum profit in case I am participant A.	5.3

A.3 Additional psychological measures

A.3.1 Big 5 Questionnaire

This questionnaire was taken from Rammstedt and John (2007).



Instruction: How well do the following statements describe your personality?

I see myself as someone who …	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
… is reserved	(1)	(2)	(3)	(4)	(5)
… is generally trusting	(1)	(2)	(3)	(4)	(5)
… tends to be lazy	(1)	(2)	(3)	(4)	(5)
… is relaxed, handles stress well	(1)	(2)	(3)	(4)	(5)
… has few artistic interests	(1)	(2)	(3)	(4)	(5)
… is outgoing, sociable	(1)	(2)	(3)	(4)	(5)
… tends to find fault with others	(1)	(2)	(3)	(4)	(5)
… does a thorough job	(1)	(2)	(3)	(4)	(5)
… gets nervous easily	(1)	(2)	(3)	(4)	(5)
… has an active imagination	(1)	(2)	(3)	(4)	(5)

A.3.2 Context (In)dependence

This questionnaire was taken from Gollwitzer (2006). The following is an English translation of the original questionnaire in German. Agreement to an item was measured on a 6 point Likert scale from "does not apply at all" to "fully applies".

Context dependence

1. My attitudes and opinions are often determined by the circumstances.
2. My behavior often depends on the people I am spending time with at that moment.
3. My decisions often depend on the temporary circumstances.

4. I behave very differently with different people.
5. My self-image depends overall on how other people perceive me.

Context independence

1. Once I have made a choice, I do not like to change it afterwards.
2. My self-image stays the same regardless of what others say about me.
3. I advocate for my own opinion regardless of the person with whom I am interacting.
4. I am the same person in different situations also.
5. My attitudes and opinions hardly change, regardless of what happens in my life.

A.3.3 Moral disengagement

This questionnaire was taken from Bandura et al. (1996). We excluded the following categories: euphemistic language, attribution of blame and dehumanization, as they did not apply to our experimental framework. The following is an English translation of the version by Rothmund (unpublished), who validated the questionnaire in German. Agreement to an item was measured on a 6-point Likert scale from "do not agree at all" to "fully agree".

1. It is alright to beat someone who badmouths your family.
2. Arriving late is better than not coming at all.
3. It does not make sense to avoid flying to go on vacation for the sake of the environment, since everybody else does it as well.
4. It is okay to tell small lies because they don't really do any harm.
5. It is alright to lie to keep your friends out of trouble.
6. Given the million-dollar frauds of some mangers, one cannot be blamed for scrounging some office supplies.
7. It is not so bad to cheat on taxes, since everybody does it anyway.
8. One cannot be blamed for an offence, if one has been put under pressure by one's friends.
9. Teasing someone does not really hurt them.
10. It is less bad to steal from the rich than from the poor.
11. A single person cannot be blamed for misbehaving, if everyone else does the same.
12. Managers cannot be blamed for layoffs, that is simply how business life works.
13. It is alright to leave some trash in the cinema hall, since it will be cleaned after the screenplay anyway.
14. The reason why poor people do not have money is that they are too lazy to work.

A.3.4 Moral identity

This questionnaire was originally developed by Aquino and Reed (2002). We used the German version validated by Rothmund Gollwitzer (unpublished) and modified the list of attributes in the instructions. The following is an English translation of the material we used. Agreement to an item was measured on a 6-point Likert scale from "do not agree at all" to "fully agree".

Instructions: Below is a list of character attributes that might describe a person. The person with these attributes could be you, but also someone else.

Fair, generous, sympathetic, nice, and benign.

Imagine a person displaying exactly these character attributes. Imagine how this person would think, feel, and act. Once you have a precise image of this person, try to answer following questions.

1. It would make me feel good to be a person who has these characteristics.
2. Being someone who has these characteristics is an important part of who I am.
3. I would be ashamed to be a person who has these characteristics.
4. Having these characteristics is not really important to me.
5. I strongly desire to have these characteristics.
6. I often wear clothes that identify me as having these characteristics.
7. The types of things I do in my spare time (e.g., hobbies) clearly identify me as having these characteristics.
8. The kinds of books and magazines that I read identify me as having these characteristics.
9. The fact that I have these characteristics is communicated to others by my membership in certain organizations.
10. I am actively involved in activities that communicate to others that I have these characteristics.

A.4 Updating example

Let us assume a signal structure as in Figure 7. When there is no externality $e = 0$, with probability $c > \frac{1}{2}$ the correct signal (negative) is sent, and with $1 - c$ the signal is wrong (positive). When the externality is $e = 1$, the situation is reversed.

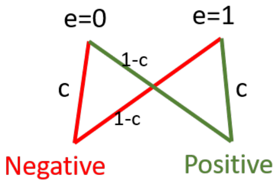


Figure 7: Examplary signal structure

The posterior given a positive or negative signal is calculated as follows (with $\hat{e}$ being the prior probability of $e = 1$). For a graphical representation, see Figure 8.

$$P_{post}(e = 1|Positive) = \frac{P(Positive|e = 1)P_{prior}(e = 1)}{P(Positive)} = \frac{c\hat{e}}{c\hat{e} + (1 - c)(1 - \hat{e})}$$
$$P_{post}(e = 1|Negative) = \frac{P(Negative|e = 1)P_{prior}(e = 1)}{P(Negative)} = \frac{(1 - c)\hat{e}}{(1 - c)\hat{e} + c(1 - \hat{e})}$$

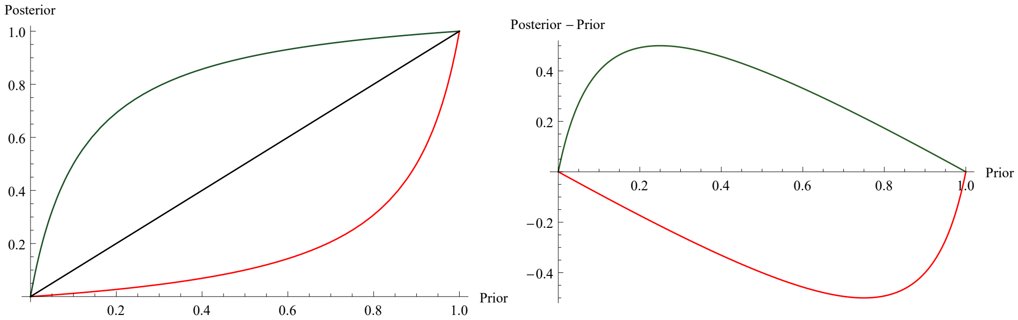


Figure 8: Posterior for given signal

Note: The left figure shows posterior beliefs as a function of prior beliefs and the right figure shows the corresponding difference between posterior and prior beliefs, both after receiving a positive (green, upper line) or negative signal (red, lower line), dependent on the prior belief. For these examples, we set $c = 0.9$. The black line on the left is the 45-degree line representing the case with no signal or no updating.

A.5 SVO measure histograms

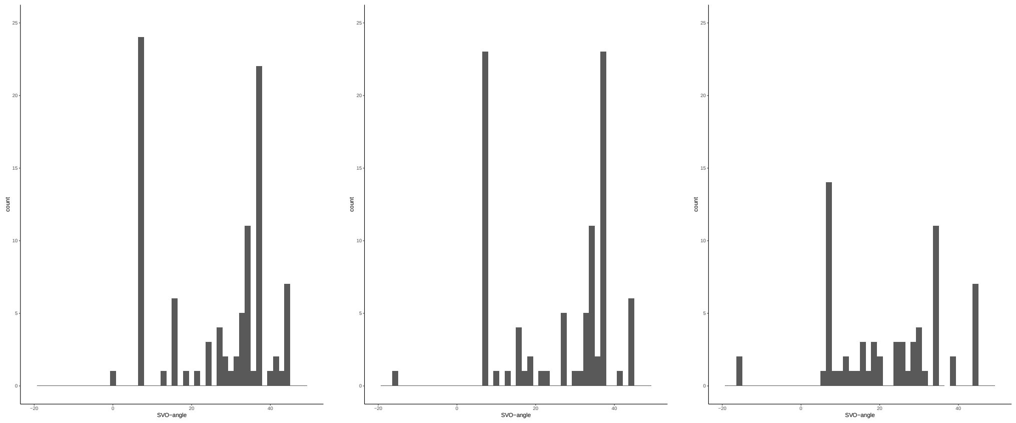


Figure 9: SVO angles in Baseline, Positive, and Negative

A.6 Robustness checks

In Table 2 we conduct multiple robustness checks. In the first column we also control for our psychological measures. In column 2, we allow for lower and upper censoring. Note that the sign and significance of treatment conditions, type, and interaction terms do not change ($p < 0.05$). For interpretability of the interactions, we plot marginal effects as in the main text (see Figure 10). Column 3 introduces a quadratic term for types and interactions with the treatment conditions (see Figure 11 for the marginal effects). We normalize our type measure for this specification (in the graph, we show the most frequent non-normalized types as references). The pattern described above remains qualitatively the same for all these alternative specifications. In column 4, we run a standard OLS regression. Coefficients have the same sign and significance level as in the Tobit regressions.

	Tobit controls	Tobit, upper and lower censoring	Tobit quadratic	OLS
POSITIVE	2.419***	5.799***	6.856	1.468***
	(2.83)	(2.68)	(1.51)	(2.64)
NEGATIVE	2.635***	5.494**	11.96***	1.313**
	(3.03)	(2.54)	(3.06)	(2.35)
Type	0.165***	0.405***	38.78***	0.123***
	(7.83)	(6.61)	(3.04)	(9.22)
POSITIVE x type	-0.0580**	-0.142**	-15.32	-0.0365*
	(-2.15)	(-1.99)	(-0.93)	(-1.95)
NEGATIVE x type	-0.0905***	-0.194***	-36.41**	-0.0487***
	(-3.29)	(-2.70)	(-2.56)	(-2.60)
Type^2			-21.78**	
			(-2.03)	
POSITIVE x type^2			8.518	
			(0.61)	
NEGATIVE x type^2			26.02**	
			(2.14)	
Constant	-3.685**	-7.972***	-12.83***	-0.509
	(-1.97)	(-4.39)	(-3.61)	(-1.28)
Controls	YES			
Observations	280	280	280	280
Pseudo/adjusted R^2	0.144	0.140	0.124	0.3647

t statistics in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

Table 2: Robustness checks

Note: OLS and tobit as described above. The type measure corresponds to the SVO angle, Positive and Negative conditions are introduced as dummies. We also include interaction terms between conditions and types. Controls include our psychological measures of context dependence, context independence, moral identity scale, and moral disengagement, as well as Big-5 measures.

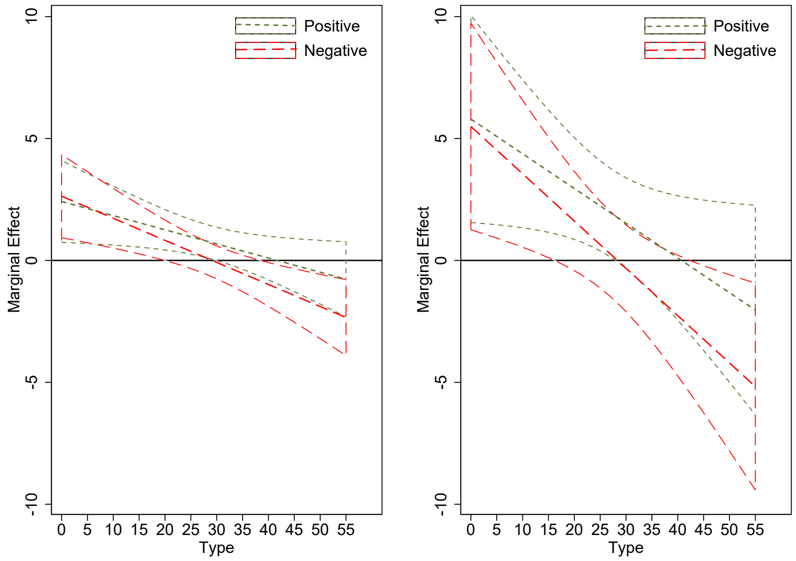


Figure 10: Marginal effects, Tobit.

Note: Tobit with controls left, Tobit with upper and lower censoring right. 95 %-confidence intervals

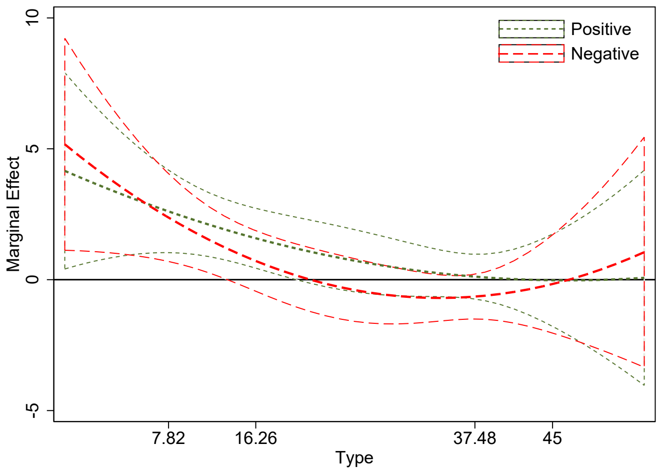


Figure 11: Marginal effects. Tobit with quadratic interaction term. 95 % confidence intervals

A.7 Analysis of additional psychological measures

In Table 3, we run the same analysis as in the Main Results section using the additional psychological measures collected in the online pre-study. Both Moral Identity and Moral Disengagement have a strong and highly significant relationship with giving in the expected direction, i.e., positive and negative respectively. However, they do not contribute to the explanation of our treatment effects. Meaning that the NEGATIVE and POSITIVE condition do not affect subjects scoring differently on these scale in a different way. As to the complementary measures of Context dependence and independence, they also do not mediate our treatment effects. Meaning that the treatment conditions do not affect subjects who are more or less dependent on the context in making their decisions, as measured by these scales, differently.

	Moral identity	Moral disengagement	Context dependence	Context independence
POSITIVE	1.500	1.705	1.485	1.391
	(0.64)	(0.88)	(1.10)	(0.64)
NEGATIVE	0.308	0.823	-0.0235	0.495
	(0.13)	(0.40)	(-0.02)	(0.23)
Measure	1.303***	-1.222**	-0.0443	0.116
	(3.25)	(-2.50)	(-0.18)	(0.28)
POSITIVE×*measure*	-0.270	-0.274	-0.211	-0.188
	(-0.48)	(-0.40)	(-0.61)	(-0.32)
NEGATIVE×*measure*	-0.133	-0.248	0.0251	-0.117
	(-0.23)	(-0.34)	(0.07)	(-0.20)
Constant	-2.913*	5.506***	2.329**	1.738
	(-1.81)	(4.08)	(2.45)	(1.13)
Observations	280	280	280	280
Pseudo R^2	0.024	0.023	0.004	0.003

t statistics in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

Table 3: Alternative measures

Note: Tobit regression with censoring at 0. Giving on treatment and stated measures as well as the interaction term.

A.8 Probit regressions

	give 5	give 0
POSITIVE	1.559**	-0.975**
	(2.39)	(-2.07)
NEGATIVE	1.020	-1.230***
	(1.47)	(-2.69)
Type	0.0820***	-0.0825***
	(4.92)	(-6.29)
POSITIVE x type	-0.0386*	0.0204
	(-1.95)	(1.09)
NEGATIVE x type	-0.0328	0.0491***
	(-1.57)	(2.95)
Constant	-2.705***	1.617***
	(-4.76)	(4.60)
Observations	280	280
Pseudo R^2	0.213	0.275

t statistics in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

Table 4: Probit regressions

Note: Probit regression. Dependent variable is a dummy of giving 5 in the first column and a dummy of giving 0 in the second column. Independent variables are treatment conditions, type, and interaction terms.

A.9 Feelings

In Table 5, we regress the measures of feelings we collected after subjects' choice in the dictator game. In all columns, we regress a specific measure on dummies for treatment conditions, the amount a subject gave, her SVO angle (type) and interaction term between the latter and the treatment conditions. The first two columns refer to general feelings of happiness and contentment (how happy/contented do you feel at the moment?). Which do not vary substantially. The last four columns refer to feelings regarding a subject's choice in the dictator game. Guilt and shame decrease in the amount a subject gives. However, the presence of narratives in our treatment conditions does not substantially alter this relationship.

	Happiness	Content	Guilt	Contentment	Shame	Excited
Constant	4.137***	3.854***	2.440***	4.169***	2.089***	2.598***
	(0.319)	(0.331)	(0.264)	(0.261)	(0.229)	(0.326)
Positive	0.694	0.756	0.455	0.318	0.240	0.553
	(0.451)	(0.468)	(0.373)	(0.369)	(0.323)	(0.461)
Negative	0.651	1.034*	−0.127	0.454	0.246	−0.027
	(0.454)	(0.470)	(0.376)	(0.371)	(0.325)	(0.464)
Type	0.013	0.017	0.012	0.018	0.005	0.001
	(0.012)	(0.013)	(0.010)	(0.010)	(0.009)	(0.013)
Give	−0.003	0.040	−0.309***	0.001	−0.213***	0.032
	(0.048)	(0.050)	(0.040)	(0.040)	(0.035)	(0.050)
Positive× Type	−0.012	−0.019	−0.014	−0.012	−0.008	−0.018
	(0.015)	(0.016)	(0.012)	(0.012)	(0.011)	(0.015)
Negative× Type	−0.017	−0.023	0.008	−0.017	−0.002	0.000
	(0.015)	(0.016)	(0.013)	(0.012)	(0.011)	(0.016)
Adj. R^2	−0.004	0.009	0.210	−0.005	0.162	−0.012
Num. obs.	280	280	280	280	280	280

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5: Regression analysis for measures of feelings

Note OLS of stated feeling on treatment, type, and the interaction term. The first two columns are general feelings, the last 4 columns are feelings specific to the choice.

A.10 Instructions

[Original instructions in German are available on request].

Welcome to the experiment

Thank you for your participation in this experiment. Please read the instructions carefully. For your participation today you will receive 5 €. During the experiment you will have the possibility to earn further money. Your additional payment will depend on your choices, the choices of other participants, as well as random events. Additionally, you will receive the earnings from the online part of the experiment at the end of today's experiment. After the experiment there will be a short questionnaire.

Please avoid any communication with your neighbors during the experiment. Switch off your mobile phone and remove everything you do not require from the table. If you have any questions, please raise your hand and we will come to answer your questions at your seat.

Instructions

In this experiment, a participant decides in the role of **Participant A** how to distribute 10 € between himself and another randomly determined **Participant B**.

First, all participants decide **in the role** of **Participant A**. This means that you will decide how to distribute **10 €** between yourself and **Participant B**. You can allocate any amount between 0 € and 10 € in discrete intervals to

Participant B. Participant B will receive this amount and you will receive the remaining amount. Your decisions will be kept anonymous and you will not know, neither during nor after the experiment, with which participant you interacted.

You will learn which role you have been assigned to only at the end of the experiment and after you have taken you decision. Half of the participants will be assigned the role of Participant A, while the other half of the participants will be assigned that of Participant B. That is, there are two possibilities:

1. You are selected as Participant A. This means: Your decision will be implemented. You will be randomly assigned to someone in the role of Participant B. You will receive 10 €, minus the amount you have allocated to Participant B. Accordingly, Participant B will receive this amount.

2. You are selected as Participant B. This means: Your decision will not be implemented. You will be randomly assigned to someone in the role of Participant A. You will receive an amount of money according to the decision of Participant A.

Since, at the time of making your your decision, you do not know whether you will be selected as Participant A or Participant B, please take your decision carefully.

After the experiment, a short questionnaire will follow. Then, the experiment will be concluded. We kindly ask you to stay seated. We will call participants individually and pay them in private. Do you have further questions? Then, please raise your hand and we will come to answer your questions at your seat. Before the actual experiment starts, you will have to answer some control questions.