The differential effect of narratives on prosocial behavior

Adrian Hillenbrand
Eugenio Verrina
The differential effect of narratives on prosocial behavior

Adrian Hillenbrand / Eugenio Verrina

December 2018,
revised September 2019
The differential effect of narratives on prosocial behavior

Adrian Hillenbrand and Eugenio Verrina*

September 11, 2019

Abstract

We study how positive narratives (stories in favor of a prosocial action) and negative narratives (stories in favor of a selfish action) influence prosocial behavior. Our main findings are that positive narratives increase giving of selfish types substantially, 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 discuss two potential explanations for this effect: one based on the enhanced saliency of normative behavior through narratives, and another based on a social comparison argument.

Keywords: Prosocial behavior, narratives, justifications, motivated moral reasoning, dictator game, SVO

JEL Classification: C91, D63, D64, D83, D91

*Hillenbrand: Max Planck Institute for Research on Collective Goods, 53113 Bonn, Germany (email: hillenbrand@coll.mpg.de). Verrina: Max Planck Institute for Research on Collective Goods, 53113 Bonn, Germany, and Cologne Graduate School in Management, Economics and Social Sciences, 50923 Cologne, Germany (email: verrina@coll.mpg.de). Financial support by the Max Planck Society is gratefully acknowledged. This experiment falls under the Generalized Approval of Standard Economic Experiment given by the Ethics Council of the Max Planck Society (Application No: 20183).
1 Introduction

Imagine that for some days you have seen a beggar on your way to work. As you pass by today, you reach into your pocket to get some change. While doing so, you remember what a colleague told you 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, according to your colleague, the beggar will spend all the money you give him on alcohol and drugs; he deserves no consideration at all. Now imagine your colleague telling you instead that rising inequality is destroying our society and that the government does not do enough for people in need. He says we should all fight against the unfairness of this wicked capitalistic system. Will you give something to the beggar after recalling one of the 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, cognitive dissonance theory (Festinger, 1962) indicates how such reasons can often be used beyond that to resolve tensions between beliefs and actions (Akerlof and Dickens, 1982). 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 stories that come in the form of rationales or justifications narratives. The notion of narratives is deeply grounded in psychological theories (Bruner, 1991; McAdams, 1988), where they serve as tools people use to construct their own account of the world. As such, narratives accompany nearly all our decisions, often playing a decisive role in shaping them. Their relevance for economic outcomes has recently received growing attention. Narratives help explain fluctuations in markets (Shiller, 2017) and also broader historical phenomena (Akerlof and Snower, 1982). Epley and Gilovich (2016) make a very similar point in their discussion of the mechanics behind motivated reasoning in general.
2016). Recent theoretical work by Bénabou et al. (2018) has contributed to the understanding of how narratives affect moral or prosocial behavior. The authors develop a model where individuals with self and social image concerns produce and consume narratives as signals complementing their actions. Unfortunately, naturally occurring data do not allow to isolate the effect of these moral arguments, since they often are bundled together with other types of information. This poses serious challenges in getting at the causal effect of narratives as rationales in favor of a certain behavior.

In this paper, we test how providing narratives affects prosocial behavior by leveraging the control of a laboratory experiment. In particular, we look at the impact of positive and negative narratives, as defined by Bénabou et al. (2018). 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, are arguments justifying 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. By controlling for the prosocial inclination of individuals, we analyze whether positive or negative narratives affect different types of individuals differently. Heterogeneity in this dimension plays an essential role in theories explaining prosocial behavior (see, e.g., Bénabou and Tirole, 2006) and recent empirical evidence confirms that individuals’ prosocial preferences greatly vary (Falk et al., 2018).

In our experiment, subjects play a dictator game where they decide how to share a given amount of money with another anonymous partici-

---

2Bénabou et al. (2018) also discuss “imperatives”, i.e., statements issued by a moral authority dictating to follow a given behavior, as an alternative way to convey moral arguments. The authors present a model, in which a principal who cares about the welfare of an agent can choose to send her either a narrative or an imperative. We focus on settings in which no such authority exists or in which she does not have enough persuasive power to issue an imperative.

3Foerster and van der Weele (2018a) work out a similar model where two agents with social image concerns can exchange signals about the social returns to an investment in a public good in a simultaneous pre-play communication phase. Their model generates a set of predictions about the use of the signals which are comparable with Bénabou et al. (2018) for what concerns the focus of this paper. In a companion paper, Foerster and van der Weele (2018b) also test this model.

4We focus on prosocial behavior as an important component of moral behavior. As opposed to prosocial behavior, we equate immoral behavior to selfish behavior.
In our two treatment conditions, they are shown either negative or positive narratives while making their choice. Narratives in the negative condition are arguments in favor of the selfish action, i.e., giving nothing to the other participant, while narratives in the positive condition are reasons in favor of the prosocial action, i.e., splitting the amount of money equally.\footnote{Krupka and Weber (2013) provide compelling empirical evidence that the equal split is indeed considered 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, while what we call the selfish action would be the strongest possible deviation from the social norm. As hinted in our behavioral predictions (see Section 3.2), our hypotheses also hold in that context.} We capitalize on arguments subjects used in previous experimental sessions for justifying their own choice to construct our treatments. This confers greater internal validity to our experimental design and allows us to systematically study the effect of the content of narratives, i.e., their appeal to the selfish or the prosocial action. We compare our two treatments to a baseline condition with no narratives. Importantly, we keep empirical expectations across all our conditions constant by showing subjects a distribution of choices made in similar dictator game experiments. This ensures that our treatment manipulations do not carry any valuable empirical information about the relative frequency of choices. We thus isolate the causal effect of narratives as providing or highlighting reasons for either the selfish or the prosocial action by exploiting the convenience of experimental control. 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 thus look at how individuals who are more or less prosocial ex ante react to the narratives we show them. We also provide a theoretical framework to illustrate how externally supplied narratives influence giving of types with different prosocial orientations and derive simple hypotheses to provide a benchmark comparison for our experimental results. According to our predictions, positive narratives should increase aggregate giving, while negative narratives should decrease it. The effect should go in the same direction for all social\footnote{We use the term “social” types to indicate all individuals with different prosocial orientations and the terms “prosocial” (or prosocials) and “selfish” to refer to individuals with high or low prosocial concerns.} types and should be stronger for prosocial types
who see a negative narrative and selfish types who see a positive narrative.

Our main results are 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 at odds with our hypotheses, according to which the same narrative cannot cause certain types to increase and other types to decrease giving. We suggest two potential explanations for this effect. According to the first, narratives - both positive and negative - enhance the salience of the moral decision, thus making it harder for subjects to behave selfishly. According to the second explanation, narratives provide a benchmark for social comparison. Since the benchmark is set at a very low level in the Negative condition, it is relatively cheap for subjects to fare better in the comparison by giving slightly more than zero.

Our results have both theoretical and practical implications. Regarding the first, they pose new challenges for theoretical models of prosocial behavior, which would need to capture the differential effect of narratives on different social types. From a practical standpoint, our results suggest that organizations and institutions wishing to promote prosocial outcomes can nudge relevant actors to confront themselves with different narratives, both positive and negative. In particular, the evidence we present indicates that such interventions would increase prosocial behavior especially among the ones who would be less inclined to act so ex ante.

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, 2018a; Shiller, 2017; Akerlof and Snower, 2016) and, more generally, in the role 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 of so-called 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; Exley, 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 evidence on how people react to externally provided narratives and by analyzing how heterogeneity in prosocial concerns affects behavior in this context.

Andreoni and Rao (2011) study a setting in which Receivers and Dictators in a dictator game can communicate with each other. They find that giving increases whenever Receivers can say something. Whereas, if only Dictators have the word, giving decreases. We investigate a setting in which Dictators are exposed to arguments coming from other Dictators, who behaved either prosocially or selfishly. People are constantly exposed to such arguments both in their professional and private life. We systematically study their effect on prosocial behavior. Similarly, Mohlin and Johannesson (2008) find a positive effect of one-way communication from the Receiver to the Dictator and also from past Receivers to Dictators. Differently from these and other studies of communication in economic games (see, e.g., Bohnet, 1999; Charness and Dufwenberg, 2006), we do not look at the effect of communication between parties involved in the game. Instead, we analyze the effect that justifications or rationales, i.e., narratives, that individuals provide for their own choice have on the behavior of other individuals facing the same decision. Thus, our setup allows us to study the causal effect of the content (positive or negative) of a narrative on prosocial behavior.

Other work has looked at how contextual factors, e.g., frames (Brañas-Garza, 2007; Dreber et al., 2013) or social information (Krupka and Weber, 2009; Gino et al., 2009; Cappelen et al., 2013, 2017), influence prosocial be-
havior. We hold these channels constant and explicitly provide reasons, or narratives, for a certain action. This links our work to studies investigating the effect of moral reminders or recommendations on behavior (see, e.g., Galbiati and Vertova (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; further work by Bott et al. (2017) uses moral appeals in letters to tax payers). 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 issued by an authority\textsuperscript{7}, i.e., the experimenter, 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 messages do not come directly from the experimenter, but are naturally occurring reasons subjects in previous sessions provide for their choices. These features of our experimental design allow us to test systematically the effect of the content of narratives, i.e., their appeal in favor of the selfish or prosocial action. Last but not least, measuring prosocial concerns allows us to look at heterogeneous effects on different social types and to test the effect of what we call negative narratives more thoroughly.\textsuperscript{8}

To achieve this goal, we use the SVO slider measure by Murphy et al. (2011) to measure social types. The SVO measure has been widely used in both psychology and economics to assess heterogeneity in individual motives in social and moral dilemmas (see Balliet et al., 2009, for a meta-study on SVO and cooperation in social dilemmas), e.g. in the public-good game (see e.g. Offerman et al., 1996). 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). Grossman and Van Der Weele (2017) study a setting where

\textsuperscript{7}The moral suasion treatments in Dal Bó and Dal Bó (2014) is very close to the notion of imperatives in Bénabou et al. (2018). In this sense, our study and the one by Dal Bó and Dal Bó (2014) can be understood as testing the effect of narratives and that of imperatives, respectively.

\textsuperscript{8}Dal Bó and Dal Bó (2014) find that messages explaining the game-theoretical prediction of zero contribution have no effect on contributions. However, baseline contributions are already quite low when they introduce this manipulation and there is hardly any room for a further decrease to take place.
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 this to be indeed an important dimension to look at, since different types display not only quantitatively, but also qualitatively different reactions.

3 Experimental Design

3.1 Setup

Our experimental design consists of two main building blocks (see Figure 1), namely an online pre-study and a laboratory experiment. The laboratory 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 €). 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.

Crucially, we fixed subjects’ empirical expectations about the distribution of giving in the dictator game. This makes sure subjects could not

---

\(^9\) Subjects received the link to the pre-study one week before the experiment and had three days to complete it.

\(^10\) 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.
take the narratives in our treatment conditions as signals about the empirical distribution of giving. Subjects in all experimental conditions were presented with a graph showing the distribution of dictator game giving in similar experiments (see Figure 6 in Appendix A.1). The graph displays data from Engel (2011) restricted to studies in which 10 units of currency were used. Subjects were told the graph displayed the distribution of choices other subjects had made in similar previous experiments.\footnote{We used the following expression: "The figure shows the frequency of choices of participants in similar experiments in percentages."} The figure displays the typical bimodal distribution with modes at 5 \euro{} and 0 \euro{} and a sizeable mass in between. While holding empirical beliefs constant across our experimental conditions, the distribution does not clearly emphasize one allocation choice over the other.

**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 which subjects in the BASELINE condition had used to explain their choices. These are our narratives (see Appendix A.2). In the POSITIVE condition, subjects saw two comments in support of the equal split (giving...
5 €), while in the NEGATIVE condition they saw two comments justifying selfish behavior (giving 0 €). Subjects were (truthfully) told that these were explanations other participants had given for their choices.\textsuperscript{12} In the next paragraph, we explain how we collected and selected the narratives to devise our treatment conditions.

**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, without any prior notice, to explain the reasoning behind their choice in the dictator game.\textsuperscript{13} We used the explanations from the BASELINE condition to build the set of narratives subjects saw in the POSITIVE and NEGATIVE condition. Three independent raters, who were blind to the research question, evaluated the narratives along several dimensions. First, they were asked whether it was possible to understand what a subject had chosen in the dictator game from his or her comment and, if so, which was the most likely choice (0, 1, 2, etc.). Raters also evaluated how convincing they perceived the narrative to be (on a 7-point Likert scale).\textsuperscript{14}

We then selected the most convincing narratives in support of giving 0 € and in support of giving 5 € (using average ratings). We excluded narratives which were particularly long or repetitive. We selected four positive and four negative narratives. Each individual in the two treatment conditions saw two randomly selected narratives (at individual level). We take these steps, 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. See Appendix A.2 for the list of selected narratives.

\textsuperscript{12}We used the following expression: "Here are two explanations (Begründungen, in German), which other participants gave for their choice."

\textsuperscript{13}The exact wording was the following. "You divided the money in the following way. You: €. Participant 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.

\textsuperscript{14}Additionally, raters evaluated the narratives with regard to their creativity, profoundness, and honesty. We do not use these measures in this study.
Type elicitation. As mentioned above, the online pre-study was conducted one week prior to the laboratory experiment to avoid contamination across the two. The purpose of our online pre-study was to measure subjects’ prosocial concerns. 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 choices, the so-called SVO angle is constructed, which represents the relative weight subjects put on the payoff of others compared to their own. Subjects with an SVO angle of 0° care only about their payoff, while those with an SVO angle of 45° weigh their payoff 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 further 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 dilemma settings (see Balliet et al., 2009, for a meta-study). Moreover, high SVO types (prosocials) have been shown to 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 these measures, see Appendix A.3). We use these measures (a) as controls in a robustness check in our regression analysis, and (b) to explore the role they 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 the questions in the order in
which they were presented to participants.\textsuperscript{15}

1. General happiness and contentment.
2. Feelings with regard to dictator game choice: happiness, guilt, content, amusement, shame, pride and 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.\textsuperscript{16} The online experiment was conducted using Qualtrics, while the laboratory experiment was programmed in zTree (Fischbacher, 2007). Subjects were recruited via Orsee (Greiner, 2015). Before the start of the laboratory experiment subjects had to answer control questions to make sure they understood the experimental instructions correctly. 282 participants (64\% female, average age 24.8 years)\textsuperscript{17} 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 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 payment 14.48 €. The online pre-study lasted between 5 and 15 minutes, while the laboratory experiment took on average 40 minutes.

3.2 Behavioral Predictions

We develop a simple theoretical framework describing how prosocial behavior is influenced by narratives and derive benchmark predictions for the effect of our treatment conditions. Our approach builds on Bénabou et al. (2018), from which we borrow some key notions. While their aim is to study a broad set of phenomena, such as the emergence of narratives and their interpretation or transmission, we focus on getting a deeper understanding of the potentially heterogeneous effects of positive and negative

\textsuperscript{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. In fact, we find no variation between treatments and a high correlation with giving. For these reasons, we do not use this measure in our analysis.

\textsuperscript{16}For an overview over all sessions, see Appendix A.9.

\textsuperscript{17}For 74 subjects, this information was not recorded.
narratives on different social types.\textsuperscript{18} This gives us a self-contained theoretical framework for which we provide an intuitive description below (the full version can be found in Appendix A.4). We first outline the reasoning leading up to our hypothesis on aggregate behavior, and then further qualify our predictions for heterogeneous social types.

We start with the notion that decision makers are more inclined to act prosocially the more the consequences of their actions benefit others or the public good (e.g., Goeree et al., 2002, and see the discussion in Bénabou and Tirole, 2006). In turn, this influences the extent to which an action is perceived as appropriate. As the literature on social norms shows, changes in what is perceived as socially appropriate reliably predict changes in behavior across several settings (Krupka and Weber, 2013).\textsuperscript{19}

Similarly, decision makers care about the deservingness of the recipient(s) of their prosocial action. In distributional choices, decision makers want to avoid giving too much to an undeserving recipient and too little to a deserving recipient (Cappelen et al., 2013). However, the true deservingness of recipients is often unknown in the real world (Cappelen et al., 2018). Likewise, the perception of what is deemed as appropriate is highly flexible and prone to self-serving interpretations (Gino et al., 2016).

Narratives in our setting are arguments or justifications targeting these perceptions of deservingness or appropriateness. A positive narrative could, for example, state that the recipient is as deserving as the dictator, because both spent the same time in the lab or because roles were assigned by a random draw. By contrast, a negative narrative might undermine the perceived appropriateness of giving, e.g., by arguing that it is not necessary to give to an anonymous recipient or that everyone else would also behave selfishly, questioning the deservingness of other participants. Importantly, these stories only need to be convincing in the sense of influencing a decision maker’s perception of the situation. If positive or negative narratives are indeed successful in changing the perception of the decision maker, they will

\textsuperscript{18}In the model by Bénabou et al. (2018), types are defined as either moral or immoral. In our setting, we look at a continuum of types, where heterogeneity stems from diverging beliefs about the appropriateness and consequences of an action.

\textsuperscript{19}The main intuitions we derive from our theoretical framework would also hold in a social norms environment with heterogeneous beliefs about the appropriateness to follow the norm, as we describe in Appendix A.4.
influence behavior. Our hypothesis on aggregate behavior follows directly.

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

We now look at how the perception, and hence the behavior, of different social types is influenced by negative and positive narratives. As mentioned above, the deservingness of a recipient and the appropriateness of giving are subject to uncertainty, and their perception can be influenced by narratives. This uncertainty leaves room for diverging perceptions. In our setting, we call decision makers who perceive a recipient to be deserving or giving appropriate “prosocial” types, and the ones who believe the opposite “selfish” types.

Consider a prosocial decision maker who hears a negative narrative undermining her perception of the recipients’ deservingness. If, as we assume above, she ascribes some truth to the narrative, her perception, and hence her behavior, will change and lead her to give less. Importantly, this effect will be greater compared to that of the same negative narrative on a selfish decision maker, who had a lower perception of the recipients’ deservingness in the first place. Vice versa, a positive narrative will have a greater effect on a selfish compared to a prosocial decision maker.

**Hypothesis 2** Positive narratives have a stronger effect on more selfish types, while negative narratives have a stronger effect on more prosocial types.

4 Results

Our dataset consists of 280 independent observations spread over three experimental conditions. In the first part of this section, we analyze the evidence regarding our main hypotheses. We then provide more additional insights on the way our treatment conditions influence behavioral results.

---

20We are agnostic about where these different perceptions come from and simply require them to influence behavior. They may be deeply grounded in a decision maker or may have formed through experience, or else a decision maker might self-servingly hold a perception which allows her to act in a certain way.

21In our experiment, we use the Social Value Orientation to measure these different perceptions. A higher (lower) SVO angle corresponds to a higher (lower) perception of deservingness or appropriateness.
4.1 Main results

Subjects in the BASELINE condition 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\textsuperscript{22}, $p = .093$). Average giving in the NEGATIVE condition (2.78 €) is virtually identical to average giving in the BASELINE condition (rank-sum test, $p = .908$).\textsuperscript{23}

![Bar chart showing average giving in different conditions](image)

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

However, the aggregate results on giving provide an incomplete picture of the data. As stated in Hypothesis 2, prosocial types should respond more strongly to the NEGATIVE treatment condition and selfish types to the POSITIVE treatment condition. Although the effect should go in the same direction for all types.

Figure 3 displays the relationship between how much a subject gave in the dictator game and her social type. Giving is, as is typical in dictator games, bounded above at 5 € with only two subjects giving 6 € and many

\textsuperscript{22}All tests are two-sided unless otherwise mentioned.

\textsuperscript{23}The difference in giving between POSITIVE and NEGATIVE is not significant (rank-sum test, $p = 0.114$).
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 the social 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 social type measure 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 many subjects giving either half of their endowment or nothing at all.

To test how different types react to different narratives, we run a Tobit regression with the amount of giving as the dependent variable and treat-
Table 1: Tobit regressions.

<table>
<thead>
<tr>
<th>dv: giving</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSITIVE</td>
<td>0.752** (0.360)</td>
<td>2.852*** (0.888)</td>
</tr>
<tr>
<td>NEGATIVE</td>
<td>0.125 (0.360)</td>
<td>2.698*** (0.894)</td>
</tr>
<tr>
<td>Type</td>
<td>0.133*** (0.0116)</td>
<td>0.189*** (0.0217)</td>
</tr>
<tr>
<td>POSITIVE x type</td>
<td>-0.0732** (0.0283)</td>
<td></td>
</tr>
<tr>
<td>NEGATIVE x type</td>
<td>-0.0900*** (0.0285)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-1.382*** (0.428)</td>
<td>-3.015*** (0.696)</td>
</tr>
</tbody>
</table>

Observations 280 280
Pseudo $R^2$ 0.108 0.118

Standard errors in parentheses
* $p < .10$, ** $p < .05$, *** $p < .01$

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 the SVO angle in column (2).

We first look at column (1), where we introduce our treatment conditions as dummies and control for the social type of a subject. The POSITIVE condition has a strong positive and significant effect on giving, confirming part of Hypothesis 1. The overall effect of the NEGATIVE condition is

---

24Due to the construction of the measure specific SVO angles appear more frequently in the data (see Murphy et al., 2011).
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. Numbers indicate the SVO-angle with higher angles indicating more prosociality. For the ease of visualization, types below 0° (3 subjects) are not displayed.

also positive, but small and not significant. Note that, as expected, the type measure is a clear predictor of giving: the higher the SVO angle of a subject, the more she gives.

In column (2) we add an interaction between subjects’ social type and the treatment conditions. To interpret these results we plot the estimated marginal effects of our treatment conditions on giving compared to the Baseline in Figure 4. This enables us to test Hypothesis 2.

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. The estimated marginal effect for the modal selfish type corresponds to a positive and significant difference of \(2.28 \text{ €} (p = .001)\) in giving, compared to the Baseline. Prosocial types, on the other hand, display no significant
increase. This finding is corroborated by comparing giving in the **Positive** condition with the **Baseline** for the modal selfish (t-test, $N = 46, p = .028$) and prosocial types (t-test, $N = 39, p = .770$) in isolation.

**Result 1 (Positive Narratives)** *Positive narratives increase giving compared to the Baseline condition. This effect is driven by more selfish types.*

In the **Negative** condition (red dashed line), more selfish types increase their giving compared to the **Baseline**. The estimated marginal difference of $2\ E (p = .004)$ for the modal selfish type is positive and significant. Note that this increase is indistinguishable from the one of the **Positive** condition. This is clearly not in line with our hypotheses. More prosocial types, on the other hand, give less than in the **Baseline**. The modal prosocial type decreases giving by an estimated $0.67\ E (p = .121)$, which is not statistically significant. However, for more prosocial types (21 subjects with an SVO angle above $44^\circ$), the effect becomes negative and significant. These results are confirmed when restricting the analysis to the modal selfish (t-test, $N = 37, p = .030$) and modal prosocial (t-test, $N = 42, p = .016$), who increase and decrease giving, respectively.\(^{25}\)

**Result 2 (Negative Narratives)** *Negative narratives have a differential effect: they decrease giving for more 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.5). 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 and session dummies as controls in our Tobit model. We also run a Tobit model with both lower and upper 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.\(^{26}\)

\(^{25}\)Note that this is in line with the LOESS fit presented in Figure 3.

\(^{26}\)We also perform our analysis using the Moral Identity Scale and the Moral Disengagement questionnaire as alternatives to the SVO angle in our main regression. Both have a strong and stable relationship with giving, but turn out to be irrelevant in explaining our treatment difference. 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.6) for further details.
4.2 Additional results

Figure 5: Marginal effects, Probit

Note: The dependent variable is a dummy for giving 5 € on the left and for giving 0 € on the right. Explanatory variables are: the SVO angle, dummies for the POSITIVE and the NEGATIVE condition and interaction terms between treatment conditions and the SVO angle. Outer lines show 95% confidence intervals. For the ease of visualization, subjects with an SVO angle below 0° (3 subjects) are not displayed.

A natural question is whether narratives led subjects to adhere to the behavioral prescription contained in them, i.e., either to share equally or keep everything for themselves. 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 regressions on the probability of giving either 5 or 0. The graphs in Figure 5 show the estimated marginal effects on different social types for the same specification we used in our main regression in Table 1 column (2) (see Table 4 in Appendix A.7 for the full regression results). There are three main observations to be made. First, the left graph in Figure 5 shows that the probability of giving an amount equal to 5 € in the POSITIVE condition increases for nearly all selfish types.\footnote{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°).} This translates into a 26% higher probability of giving 5 € for the modal selfish type in the POSITIVE condition (estimated marginal effect, \( p = .022 \)). In the NEGATIVE condition, on the contrary, 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, the right graph in Figure 5 shows that both in the POSITIVE and the NEGATIVE condition the probability of selfish types giving 0 decreases substantially. This effect is observed across a wider range in the POSITIVE condition. The estimated marginal decrease in the probability of giving 0 € for the modal selfish type 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 lead to a substantial increase in the probability of giving 0 €. The increase in probability for the modal prosocial type is moderate (11\%) and only marginally significant \((p = .077)\).

**Result 3** The POSITIVE condition increases the probability of giving 5 € for selfish types. Both treatment conditions decrease the probability of selfish types giving 0 €. The NEGATIVE condition leads to a small 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.8 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 effects is caused by the anticipation of these feelings. The presence of narratives could lead subjects to anticipate guilt or shame and to adapt their giving to avoid them, which could result in similar stated feelings across treatments.

**Result 4** Our treatment conditions do not directly change subjects’ feelings towards their choice.

5 Discussion and Conclusion

Our results provide 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 our two treatment
conditions with a baseline in which no narratives are provided. Empirical
beliefs about the distribution of choices are fixed across all experimental
conditions. We work out two hypotheses from a theoretical framework on
how narratives influence behavior via the perception of the appropriateness
of an action or the deservingness of a recipient 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 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 the selfish types in the Positive
condition. These results are only partly in line with the hypotheses derived
from our theoretical framework. In particular, our hypotheses allow the
effect of narratives to have different strength for different social types, but
predict it should go in the same direction for all social types.

The differential effect of narratives resonates well with other research
which shows 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, the process
underlying prosocial and selfish decisions is 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 explanations which can account for the dif-
ferential effect: one based on the argument that narratives enhance the
moral saliency of the decision and another one based on a social compari-
son argument. According to the first explanation, as pointed out above, the
more selfish individuals might not consider the moral consequences of their
actions in their “ordinary” decision process. They could genuinely not be
aware of these consequences or self-servingly suppress them. In both cases,
the mere presence of a narrative, regardless of its content, could make the
moral nature of the situation more salient, leading selfish individuals to
give more. This conjecture is in line with a study by Krupka and Weber (2009), who find that descriptive information enhances prosocial behavior, even in cases where one 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. To the extent that seeing a narrative activates the need to justify one’s own behavior towards oneself, our findings are also coherent with a study by Xiao (2017), who shows that the pressure to justify leads to more norm-compliant behavior in prosocial choices. Finally, Lazear et al. (2012) demonstrate that a substantial part of sharing in dictator games comes from “reluctant sharers”, who would prefer to avoid sharing, but conform to the norm to avoid feeling guilty or ashamed, or to escape social pressure. In this sense, narratives might activate such mechanisms, thus making giving unavoidable.

The second explanation relies on a subject comparing herself with the narrator. If subjects care about how they fare in the comparison with others, 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, they can decrease their giving slightly, while withholding a positive image of themselves. This account is coherent with psychological theories which emphasize the importance of social comparison for people’s self-perception (Festinger, 1954) and with empirical evidence showing that people indeed adapt their behavior accordingly (Frey and Meier, 2004).

Our study 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. The observed increase of equal splits in the positive condition can be interpreted as an effect of the enhanced moral salience, which pushes subjects to comply with the norm of behavior more often. On the other hand, the fact that in the negative condition both the probability of prosocial types giving nothing and that of selfish types splitting equally do not substantially increase seems to point towards
a more gradual tradeoff, which might be the result of social comparison. The behavior of prosocial types is compatible with the phenomenon of partial lying (Fischbacher and Föllmi-Heusi, 2013) or ethical maneuvering (Mazar et al., 2008; Shalvi et al., 2011b), which is consistently found in the experimental literature on lying and cheating. Subjects often 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 in our experiment 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. Hence, narratives in favor of the moral action, it appears, lead people to comply with their behavioral prescription, instead, in the case of negative narratives, people engage in a more gradual trade-off.

Our work advances the understanding of the determinants of prosocial and moral behavior. Narratives permeate every aspect of human behavior and are fundamental tools guiding people’s actions. Our study provides causal evidence showing that these factors can have a sizeable impact on economic behavior. The presence of a narrative seems to lead individuals to think more about their choice, which in turn pushes them to act more prosocially. This has relevant implications for institutions or organizations that wish to promote prosocial behavior. Nudging actors to confront themselves with arguments in favor or against normative behavior can increase compliance, especially amongst the ones who would be less inclined to act so ex ante.

In the setting we study, sharing the money equally represents a clear norm of behavior. Future research should investigate the relationship between narratives and the strength of a norm or the presence of multiple norms. Other questions are how enduring the effect of a certain narrative is, and whether there might be spillovers in other contexts. From a theoretical point of view, a model with the aim of accounting for the evidence presented here would need to capture the differential effect of narratives on different social types. We hope our work can contribute to inspire such endeavors.
References


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


_ and _, “Persuasion, justification and the communication of social impact,” 2018.


_, Jason Dana, Michel JJ Handgraaf, and Carsten KW De Dreu, “Justified ethicality: Observing desired counterfactuals modifies ethical


A Appendix

A.1 Decision Screen

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. Below subject take the dictator game decision.

A.2 Narrative Selection

The following table shows positive and negative narratives (translated from German) 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. More detailed information as well as the complete list of comments is available from the authors upon request.
<table>
<thead>
<tr>
<th>Number</th>
<th>Positive Narratives</th>
<th>Convincingness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Both came here to participate in the experiment and spent the same amount of time here. Both should get the same payment.</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>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.</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>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.</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>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.</td>
<td>6.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative Narratives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>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.</td>
</tr>
<tr>
<td>6</td>
<td>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 role B.</td>
</tr>
<tr>
<td>7</td>
<td>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.</td>
</tr>
<tr>
<td>8</td>
<td>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.</td>
</tr>
</tbody>
</table>

A.3 Additional psychological measures

A.3.1 Big 5 Questionnaire

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

<table>
<thead>
<tr>
<th>Instruction: How well do the following statements describe your personality?</th>
</tr>
</thead>
<tbody>
<tr>
<td>I see myself as someone who ...</td>
</tr>
<tr>
<td>… is reserved</td>
</tr>
<tr>
<td>… is generally trusting</td>
</tr>
<tr>
<td>… tends to be lazy</td>
</tr>
<tr>
<td>… is relaxed, handles stress well</td>
</tr>
<tr>
<td>… has few artistic interests</td>
</tr>
<tr>
<td>… is outgoing, sociable</td>
</tr>
<tr>
<td>… tends to find fault with others</td>
</tr>
<tr>
<td>… does a thorough job</td>
</tr>
<tr>
<td>… gets nervous easily</td>
</tr>
<tr>
<td>… has an active imagination</td>
</tr>
</tbody>
</table>

A.3.2 Context (In)dependence

This questionnaire is taken from Gollwitzer (2006). The following is an English translation of the original questionnaire in German. Agreement to an item is 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.
5. My attitudes and opinions hardly change, regardless of what happens in my life.

A.3.3 Moral disengagement

This questionnaire is 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 managers, 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 he or she has been put under pressure by his or her 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 use 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 is 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 conveyed to others by my
   membership in certain organizations.
10. I am actively involved in activities that convey to others that I have these
characteristics.

A.4 Theoretical framework

This section complements the “Behavioral Predictions” in the main text (Section
3.2) by providing formal definitions and derivations of the hypotheses. A decision
maker chooses how much money to give to a recipient. A key component of this
model is the belief about the externality of giving (Bénabou et al., 2018). We,
first, describe the basic utility function of a decision maker; we, then, explain
which role the externality plays; and, finally, discuss how narratives enter the
model.

The utility function of a decision maker (DM) takes the following form:

\[ U_i(g, e) = v(g, e) - c(g), \]

where \( g \) is the amount she decides to give, and \( e \) is the expected externality of
giving, which we define below; \( v(g, e) \) captures the overall valuation of giving, and
We set $e \in (0, 1)$ and assume $c(g)$ to be linear increasing 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 externality.** $E$ is a binary measure of the presence of a positive externality, i.e., whether the recipient is deserving or it is appropriate to give in the situation at hand (see discussion in Section 3.2). If $E = 1$, there is a positive externality, while if $E = 0$, there is no such externality. A DM in our model does not know the value of $E$ with certainty. Rather, he holds a prior belief (what we call perception above) about $E$ with $e = P(E = 1)$. We assume that the marginal utility of giving is increasing in the expected externality $e$ ($\frac{\partial v(g, e)}{\partial g} > 0$). 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 understood as a norm. Correspondingly, by setting $v(g, e) = -\gamma(e)(\frac{1}{2} - g)^2$ in a dictator game with a pie size of 1, $\gamma(e)$ would capture the appropriateness to follow the norm, i.e., to split the pie equally (assuming $\frac{\partial \gamma}{\partial e} > 0$). Independently of the specific choice of $v$, our predictions hold.

**Narratives.** We model narratives as signals about $E$ updating the prior belief of a DM, as in Bénabou et al. (2018). A positive narrative signals that $E = 1$, i.e., it is an argument or justification for there being a positive externality. A negative narrative, conversely, signals that $E = 0$. For simplicity, we take DMs to be standard Bayesian updaters. Other forms of updating are of course conceivable, but would introduce further degrees of freedom in the model. 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. We assume narratives to be at least somewhat believable or convincing, which here means that the signal is correct more often than not. Hence, a DM will update in the direction of the signal.\footnote{Note that Bénabou et al. (2018) formally define positive and negative narratives directly by their influence on beliefs. The signalling structure we use is based on an older version of their paper and leads to the same directional effect of narratives on actions.}

As an example, let us assume a signal structure as in Figure 7. If there is

\footnote{Note that all factors influencing the utility of giving are captured by the first term. For the sake of simplicity, we do not consider how image concerns would alter the resulting trade-off.}
no externality \( E = 0 \), with probability \( 1 \geq c > \frac{1}{2} \) the correct signal, i.e. the negative narrative, is sent, and with \( 1 - c \) the signal is wrong, i.e. the narrative is positive. The situation is reversed with a high externality \((E = 1)\).

Figure 7: Exemplary signal structure

The posterior given a positive or negative signal is calculated as follows (with \( e \) being the prior probability of \( E = 1 \)). Figure 8 provides a graphical representation.

\[
\begin{align*}
P_{post}(E = 1|\text{Positive}) &= \frac{P(\text{Positive}|E = 1)P_{prior}(E = 1)}{P(\text{Positive})} = \frac{ce}{ce + (1 - e)(1 - e)} \\
P_{post}(E = 1|\text{Negative}) &= \frac{P(\text{Negative}|E = 1)P_{prior}(E = 1)}{P(\text{Negative})} = \frac{(1 - c)e}{(1 - c)e + c(1 - e)}
\end{align*}
\]

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.

Given this signal structure, negative narratives lead to a downward shift in beliefs and positive narratives to an upward shift. 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.*

**Heterogeneity.** We introduce heterogeneity by allowing diverging beliefs about $E$. In fact, DMs in our model differ solely in their beliefs, which we bound to $e \in (0, 1)$. That is, all DMs in our model would act in the same way, i.e., choose to give the same amount, if they held the same belief. Modelling heterogeneity solely through beliefs offers us a concise way to introduce narratives as signals. We call DMs with low beliefs “selfish” types and those with high beliefs “prosocial” types.

While in our framework the direction of the effect of narratives is independent from prior beliefs, our setup predicts a different strength of the effect for different priors. 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 8).

**Hypothesis 2** *Positive narratives should have a stronger positive effect on more selfish types, while negative narratives should have a stronger negative effect on more prosocial types.*

---

30Bénabou et al. (2018) hint at heterogeneity in priors, but consider common priors throughout the paper with heterogeneity between subjects stemming solely from different valuations of the externality.
A.5 Robustness checks

In Table 2 we conduct multiple robustness checks. In the first column we control for the additional psychological measures. In column 2, we impose both lower and upper censoring. For interpretability of the interactions, we plot marginal effects as in the main text (see Figure 9). Column 3 introduces a quadratic term for types and interactions with the treatment conditions (see Figure 10 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 in Section 4 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.

Figure 9: Marginal effects, Tobit.

*Note:* Tobit with controls left, Tobit with upper and lower censoring right. 95 % confidence intervals
<table>
<thead>
<tr>
<th></th>
<th>Tobit controls</th>
<th>Tobit sessions</th>
<th>Tobit, upper and lower censoring</th>
<th>Tobit quadratic</th>
<th>OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POSITIVE</strong></td>
<td>2.419***</td>
<td>1.884*</td>
<td>5.799***</td>
<td>6.856</td>
<td>1.468***</td>
</tr>
<tr>
<td></td>
<td>(0.855)</td>
<td>(1.062)</td>
<td>(2.166)</td>
<td>(4.548)</td>
<td>(0.555)</td>
</tr>
<tr>
<td><strong>NEGATIVE</strong></td>
<td>2.635***</td>
<td>2.709**</td>
<td>5.494**</td>
<td>11.96**</td>
<td>1.313***</td>
</tr>
<tr>
<td></td>
<td>(0.868)</td>
<td>(1.086)</td>
<td>(2.162)</td>
<td>(3.907)</td>
<td>(0.560)</td>
</tr>
<tr>
<td>Type</td>
<td>0.165****</td>
<td>0.163***</td>
<td>0.405**</td>
<td>38.78***</td>
<td>0.123***</td>
</tr>
<tr>
<td></td>
<td>(0.0211)</td>
<td>(0.0211)</td>
<td>(0.0613)</td>
<td>(12.77)</td>
<td>(0.0133)</td>
</tr>
<tr>
<td><strong>POSITIVE x type</strong></td>
<td>-0.0580**</td>
<td>-0.0532**</td>
<td>-0.142**</td>
<td>-15.32</td>
<td>-0.0365**</td>
</tr>
<tr>
<td></td>
<td>(0.0270)</td>
<td>(0.0269)</td>
<td>(0.0717)</td>
<td>(16.54)</td>
<td>(0.0187)</td>
</tr>
<tr>
<td><strong>NEGATIVE x type</strong></td>
<td>-0.0905***</td>
<td>-0.0918***</td>
<td>-0.194***</td>
<td>-36.41**</td>
<td>-0.0487***</td>
</tr>
<tr>
<td></td>
<td>(0.0275)</td>
<td>(0.0275)</td>
<td>(0.0717)</td>
<td>(14.22)</td>
<td>(0.0188)</td>
</tr>
<tr>
<td><strong>Type^2</strong></td>
<td></td>
<td></td>
<td></td>
<td>-21.78**</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(10.74)</td>
<td></td>
</tr>
<tr>
<td><strong>POSITIVE x type^2</strong></td>
<td></td>
<td></td>
<td></td>
<td>8.518</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(13.99)</td>
<td></td>
</tr>
<tr>
<td><strong>NEGATIVE x type^2</strong></td>
<td></td>
<td></td>
<td></td>
<td>26.02**</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(12.16)</td>
<td></td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>-3.685**</td>
<td>-3.5625*</td>
<td>-7.972***</td>
<td>-12.83***</td>
<td>-0.509</td>
</tr>
<tr>
<td></td>
<td>(1.867)</td>
<td>(1.9015)</td>
<td>(1.815)</td>
<td>(3.558)</td>
<td>(0.397)</td>
</tr>
<tr>
<td>Controls</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Session</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Observations</td>
<td>280</td>
<td>280</td>
<td>280</td>
<td>280</td>
<td>280</td>
</tr>
<tr>
<td>Pseudo $R^2$</td>
<td>0.144</td>
<td>0.1512</td>
<td>0.140</td>
<td>0.124</td>
<td>0.3647</td>
</tr>
</tbody>
</table>

Standard errors in parentheses
* $p < .10$, ** $p < .05$, *** $p < .01$

Table 2: Robustness checks

*Note:* OLS and Tobit. 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 Context Dependence, Context Independence, Moral Identity Scale, Moral Disengagement, and the 11-item, Big-5 questionnaire. Session includes session dummies.
A.6 Analysis of additional psychological measures

In Table 3, we run the same analysis as in Section 4 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 significantly 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 do not significantly mediate our treatment effects. Meaning that the treatment conditions do not affect subjects who are more or less dependent from the context in making their decisions, as measured by these scales, differently.
### Table 3: Alternative measures

<table>
<thead>
<tr>
<th></th>
<th>Moral identity</th>
<th>Moral disengagement</th>
<th>Context dependence</th>
<th>Context independence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POSITIVE</strong></td>
<td>1.500</td>
<td>1.705</td>
<td>1.485</td>
<td>1.391</td>
</tr>
<tr>
<td>measure</td>
<td>1.303***</td>
<td>-1.222**</td>
<td>-0.0443</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td>(2.340)</td>
<td>(1.933)</td>
<td>(1.352)</td>
<td>(2.185)</td>
</tr>
<tr>
<td><strong>NEGATIVE</strong></td>
<td>0.308</td>
<td>0.823</td>
<td>-0.0235</td>
<td>0.495</td>
</tr>
<tr>
<td>measure</td>
<td>-0.270</td>
<td>-0.274</td>
<td>-0.211</td>
<td>-0.188</td>
</tr>
<tr>
<td></td>
<td>(0.567)</td>
<td>(0.676)</td>
<td>(0.344)</td>
<td>(0.583)</td>
</tr>
<tr>
<td><strong>POSITIVE × measure</strong></td>
<td>-0.270</td>
<td>-0.274</td>
<td>-0.211</td>
<td>-0.188</td>
</tr>
<tr>
<td></td>
<td>(0.567)</td>
<td>(0.676)</td>
<td>(0.344)</td>
<td>(0.583)</td>
</tr>
<tr>
<td><strong>NEGATIVE × measure</strong></td>
<td>-0.133</td>
<td>-0.248</td>
<td>0.0251</td>
<td>-0.117</td>
</tr>
<tr>
<td></td>
<td>(0.581)</td>
<td>(0.735)</td>
<td>(0.352)</td>
<td>(0.587)</td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>-2.913*</td>
<td>5.506***</td>
<td>2.329**</td>
<td>1.738</td>
</tr>
<tr>
<td></td>
<td>(1.613)</td>
<td>(1.349)</td>
<td>(0.952)</td>
<td>(1.538)</td>
</tr>
</tbody>
</table>

Observations: 280
Pseudo $R^2$: 0.024

Standard errors in parentheses

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

Note: Tobit regression with censoring at 0. Giving on treatment and stated measures as well as the interaction term.
A.7 Probit regressions

<table>
<thead>
<tr>
<th></th>
<th>give 5</th>
<th>give 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSITIVE</td>
<td>1.559**</td>
<td>-0.975***</td>
</tr>
<tr>
<td></td>
<td>(0.653)</td>
<td>(0.471)</td>
</tr>
<tr>
<td>Type</td>
<td>0.0820***</td>
<td>-0.0825***</td>
</tr>
<tr>
<td></td>
<td>(0.0167)</td>
<td>(0.0131)</td>
</tr>
<tr>
<td>POSITIVE x type</td>
<td>-0.0386*</td>
<td>0.0204</td>
</tr>
<tr>
<td></td>
<td>(0.0198)</td>
<td>(0.0186)</td>
</tr>
<tr>
<td>NEGATIVE</td>
<td>1.020</td>
<td>-1.230***</td>
</tr>
<tr>
<td></td>
<td>(0.694)</td>
<td>(0.457)</td>
</tr>
<tr>
<td>NEGATIVE x type</td>
<td>-0.0328</td>
<td>0.0491***</td>
</tr>
<tr>
<td></td>
<td>(0.0209)</td>
<td>(0.0166)</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.705***</td>
<td>1.617***</td>
</tr>
<tr>
<td></td>
<td>(0.568)</td>
<td>(0.352)</td>
</tr>
<tr>
<td>Observations</td>
<td>280</td>
<td>280</td>
</tr>
<tr>
<td>Pseudo $R^2$</td>
<td>0.213</td>
<td>0.275</td>
</tr>
</tbody>
</table>

Standard errors 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.8 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 and an 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 are rather stable. 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.
### Table 5: Regression analysis for measures of feelings

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

<table>
<thead>
<tr>
<th></th>
<th>Happiness</th>
<th>Content</th>
<th>Guilt</th>
<th>Contentment</th>
<th>Shame</th>
<th>Excited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>4.137***</td>
<td>3.854***</td>
<td>2.440***</td>
<td>4.169***</td>
<td>2.089***</td>
<td>2.598***</td>
</tr>
<tr>
<td></td>
<td>(0.319)</td>
<td>(0.331)</td>
<td>(0.264)</td>
<td>(0.261)</td>
<td>(0.229)</td>
<td>(0.326)</td>
</tr>
<tr>
<td>Positive</td>
<td>0.694</td>
<td>0.756</td>
<td>0.455</td>
<td>0.318</td>
<td>0.240</td>
<td>0.553</td>
</tr>
<tr>
<td></td>
<td>(0.451)</td>
<td>(0.468)</td>
<td>(0.373)</td>
<td>(0.369)</td>
<td>(0.323)</td>
<td>(0.461)</td>
</tr>
<tr>
<td>Negative</td>
<td>0.013</td>
<td>0.012</td>
<td>0.008</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.013)</td>
<td>(0.010)</td>
<td>(0.010)</td>
<td>(0.009)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>Type</td>
<td>-0.003</td>
<td>-0.003</td>
<td>-0.003</td>
<td>-0.003</td>
<td>-0.003</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.048)</td>
<td>(0.050)</td>
<td>(0.040)</td>
<td>(0.040)</td>
<td>(0.035)</td>
<td>(0.050)</td>
</tr>
<tr>
<td>Positive$\times$ Type</td>
<td>-0.012</td>
<td>-0.012</td>
<td>-0.012</td>
<td>-0.012</td>
<td>-0.012</td>
<td>-0.012</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.016)</td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.011)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>Negative$\times$ Type</td>
<td>-0.017</td>
<td>-0.017</td>
<td>-0.017</td>
<td>-0.017</td>
<td>-0.017</td>
<td>-0.017</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
<td>(0.016)</td>
<td>(0.013)</td>
<td>(0.012)</td>
<td>(0.011)</td>
<td>(0.016)</td>
</tr>
</tbody>
</table>

Adj. R²               -0.004   0.009  0.210  -0.005  0.162  -0.012

Num. obs.             280    280    280    280    280    280

Standard errors in parentheses

***p < 0.001, **p < 0.01, *p < 0.05

### Table 6: Session overview

<table>
<thead>
<tr>
<th>Session</th>
<th>Date (2018)</th>
<th>Treatment</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>May, 7</td>
<td>BASELINE</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>May, 16</td>
<td>BASELINE</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>May, 16</td>
<td>BASELINE</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>May, 30</td>
<td>POSITIVE</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>May, 30</td>
<td>NEGATIVE</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>May, 30</td>
<td>POSITIVE</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>May, 30</td>
<td>NEGATIVE</td>
<td>26</td>
</tr>
<tr>
<td>8</td>
<td>June, 26</td>
<td>POSITIVE</td>
<td>24</td>
</tr>
<tr>
<td>9</td>
<td>June, 26</td>
<td>BASELINE</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>June, 26</td>
<td>NEGATIVE</td>
<td>25</td>
</tr>
<tr>
<td>11</td>
<td>June, 26</td>
<td>NEGATIVE</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>June, 26</td>
<td>POSITIVE</td>
<td>18</td>
</tr>
</tbody>
</table>

A.9 Sessions
A.10 Instructions

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 need for the experiment 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. 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 your 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 the amount you allocated him.

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 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?
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 questions of understanding.