Moral awareness polarizes people's fairness judgments*

Michael Kurschilgen

November 2021

Abstract. How does moral awareness affect people's fairness judgments? Models of identity utility predict that if individuals differ in their personal fairness ideals (equality versus efficiency), higher moral awareness should not only make people's choices less selfish but also more polarized. On the other hand, people's desire for conforming with the behavior of their peers should help mitigate polarization. I test these conjectures in a laboratory experiment, in which participants can pursue different fairness ideals. I exogenously vary (i) whether participants are prompted to state their moral opinions behind the veil of ignorance, and (ii) whether they are informed about the behavior of their peers. I find that moral introspection makes choices more polarized, reflecting even more divergent moral opinions. The increase in polarization coincides largely with a widening of revealed gender differences as introspection makes men's choices more welfarist and women's more egalitarian. Disclosing the descriptive norm of the situation is not capable of mitigating the polarization.

Keywords: Moral Introspection; Social Information; Identity; Normative Ambivalence; Equality; Efficiency; Polarization; Experiment. **JEL codes:** C91, D63

* M. Kurschilgen is affiliated to the Technical University of Munich and the Max Planck Institute for Research on Collective Goods (m.kurschilgen@tum.de). Correspondence address: Michael Kurschilgen, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany. I am grateful to Samuel Bowles, Marco Casari, Christoph Engel, Werner Güth, Yuval Feldman, Shachar Kariv, Oliver Kirchkamp, Erin Krupka, Isabel Marcin, Moti Michaeli, Marta Serra-Garcia, Shyam Sunder, and seminar participants at Bar-Ilan, Bologna, Bonn, Jena, Lausanne, Tilburg and Michigan, for valuable impulses and comments on earlier versions of the paper, as well as Nicolas Meier, Carlos Kurschilgen, Caroline Roeger, and Bernd Engelmann for research assistance. This study was funded by the Max Planck Society for the Advancement of Science (MPG). The author declares that he has no conflict of interest.

1 Introduction

A large strand of research shows that raising people's moral awareness can have beneficial effects by making choices less selfish (Dal Bó and Dal Bó 2014; Schram and Charness 2015; Kessler and Milkman 2018), and more honest (Gino et al. 2011; Kouchaki and Smith 2014; Welsh and Ordóñez 2014; Feldman and Halali 2017).[1] Yet, little is known about how moral awareness affects behavior in normatively ambivalent situations, i.e. situations in which different people may have conflicting opinions of what is the right thing to do (e.g. due to diverse cultural, religious, or socio-demographic backgrounds).[2] Allocation problems – wage negotiations, distributions of bonuses, questions of fair taxation, issues of political representation – are prime examples of normatively ambivalent situations. In a recent, representative survey, Müller and Renes (2021) document the heterogeneity of fairness judgments, particularly along gender lines, with men prioritizing efficiency whilst women favor equality.[3]

If people derive utility both from material consumption opportunities and from adhering to their personal normative ideal (Akerlof and Kranton 2000; Burks and Krupka 2012; Bašić and Verrina 2020), an increase in moral awareness should make behavior not only less selfish but also more polarized. Higher awareness draws people's choices closer to their respective ideals and further away from material selfishness as the common denominator. Polarization is socially undesirable. It undermines the ability of social groups to solve cooperation problems (Dimant 2021). Countries with high levels of opinion polarization on morally charged issues tend to have lower levels of trust between individuals (Rapp 2016). In a polarized political climate, people may be willing to sacrifice democratic principles for partisan interests (Graham and Svolik 2020; Svolik et al. 2020). Within an organization, polarization may hurt productivity (Burks and Krupka 2012).

Yet, in reality, moral judgments do not occur in a vacuum. People will typically contrast their privately-held moral rules with the behavior they commonly observe from their peers. Drawing on people's desire for conforming with the behavior of their peers (Bernheim 1994; Michaeli and Spiro 2015), polarization could be mitigated by providing information about what people commonly do (Bicchieri and Dimant 2019). Potentially, social information may not only homogenize people's behavior but even their moral opinions. Studying situations without normative ambivalence, Lindström et al. (2018) report evidence suggesting that the commonness

[1] In recent years, an increasing number of companies has been encouraging their employees to actively think about the moral implications of their behavior (Kaptein 2015). Coca-Cola (2018) for instance urges employees to "do what is right". Google (2018) appeals to "identifying the right thing to do", acknowledging that it is "impossible to spell out every possible ethical scenario we might face. Instead, we rely on one another's good judgment [...]".

[2] A notable exception is the theoretical work of te Velde (2020).

[3] More broadly, conceptions of morality have been shown to differ between liberals and conservatives (Haidt and Graham 2007; Graham et al. 2009), men and women (Friesdorf et al. 2015), and across cultures (Graham et al. 2016).

of behavior influences people's moral judgments. On the other hand, information disclosure could have the inverse effect if it mainly directed people's focus toward the discrepancy between what they observe and what they believe to be right (Cialdini et al. 1990; Krupka and Weber 2009), thereby reinforcing the weight of the latter.

Testing the above conjectures requires (i) a choice environment in which people may plausibly pursue different moral ideals, (ii) a measure of people's underlying moral ideals, and (iii) exogenous variation of moral awareness and social information. This paper reports evidence from a laboratory experiment with those exact features. Specifically, I conduct a modified dictator game (MDG), in which participants face (a) a trade-off between their material benefit and their personal distributional ideal, and (b) a trade-off between the ideal of "equality" and the ideal of "efficiency" (Fisman et al. 2007; Iriberri and Rey-Biel 2013). The MDG consists of 2×2 decision panels, in which the dictator (a) is either richer or poorer than the recipient, and (b) can sacrifice own payoffs to either destroy or create recipient's payoffs, at different relative prices. This game has an unambiguous selfish optimum but several plausible moral optima, ranging from maximizing total welfare, over Rawlsian Maximin, to minimizing inequality.

Between treatments, I exogenously vary (i) whether there is an additional stage in which participants are asked for their moral opinions *before* they know whether they will be dictators or recipients in the subsequent MDG, i.e. behind the veil of ignorance, and (ii) whether participants are informed about the behavior of their peers. For each decision task, subjects are shown which option was chosen by the majority of previous dictators and how large the respective majority was.[4]

My experimental results show that subjects who were exogenously assigned to reflect on what is "morally right" before the game have a 92% higher willingness to sacrifice monetary payoffs in the MDG. Yet, payoffs are sacrificed for rather different causes. While some participants invest them into increasing total surplus, even if this implies higher inequality, others invest into increasing equality, even if this implies lower total surplus. As a result, raising moral awareness increases the polarization of choices by 33%. In fact, participants' answers to the question "what do you find morally right" are even more polarized than their subsequent, incentivized choices, revealing pronounced home-grown discrepancies in participants' conceptions of morality.

In contrast, I find no significant effect of social information, neither on the level of selfishness, nor on the level of polarization. This is striking since the social information showed the – predominantly selfish – true behavior of participants in the *Baseline*. In principle, sub-

[4] Following the taxonomy of Bicchieri and Dimant (2019), my two treatment dimensions represent two dichotomous interventions into the choice architecture. Whereas moral introspection aims at influencing behavior by appealing to people's *unconditional* preferences for adhering to an *injunctive* norm, providing social information seeks to affect behavior by appealing to people's *conditional* preferences for complying with a *descriptive* norm. By independently quantifying the effectiveness of both interventions, we can therefore diagnose the nature of the targeted behavior (Bicchieri 2016; Bicchieri and Dimant 2019).

jects could have used the information opportunistically to justify more selfish (and thus more homogeneous) choices. On the other hand, moral introspection has a similarly strong effect on informed participants as on uninformed ones. The willingness to sacrifice payoffs increases by 86% and polarization by 32%. Moreover, I find that the increase in polarization coincides largely with a widening of revealed gender differences as raising moral awareness makes men's choices more welfarist and women's more egalitarian.

My findings corroborate the importance of self-image or identity (Akerlof and Kranton 2000; Bénabou and Tirole 2011) for understanding people's revealed concern for others. Experimental evidence has shown that people's revealed pro-sociality drops as one's moral wiggle room (Dana et al. 2007) increases. By opportunistically picking a certain narrative (Bénabou et al. 2019), people can act selfishly while maintaining a positive self-image. People's moral wiggle room has been shown to increase by delegating responsibility to other people (Bartling and Fischbacher 2011; Hamman et al. 2010) or to market forces (Bartling et al. 2014; Falk and Szech 2013), as well as by distorting beliefs about others' likely behavior (Di Tella et al. 2015). My findings show that moral introspection can be an effective tool to reduce one's wiggle room, even in the face of social information providing an opportunistic narrative.

I thus contribute to the experimental literature on behavioral nudges aimed at restraining human selfishness (see Bicchieri and Dimant (2019) and Engel and Kurschilgen (2020) for recent reviews). Frey and Meier (2004), Bicchieri and Xiao (2009) and Shang and Croson (2009) show that people behave more pro-socially when they observe pro-social behavior of others. Krupka and Weber (2009) report that even participants who observed more selfish behavior acted more socially than participants in a baseline without information. The latter is not consistent with conformity but rather with the idea that observing anti-social behavior increases the focus on the pro-social norm (Cialdini et al. 1990). Extant work has looked into situations with a one-dimensional moral action space, in which there are clear morally superior, *social* choices and morally inferior, *selfish* choices. In the normatively ambivalent situation studied in this paper, in which subjects need to choose not only between selfish and social but also between divergent conceptions of social, the effect of social information is rather limited whereas the effect of moral introspection is substantial. This implies that nudges aimed at turning egalitarians into welfarists (or vice-versa) are not likely to be effective.

In particular, my findings add an interesting new insight to the literature on gender differences. Several survey studies report that women care more about equality and tend to be more favorable to redistribution than men (Alesina and Ferrara 2005; Corneo and Grüner 2002; Fong 2001; Ravallion and Lokshin 2000; Müller and Renes 2021). This has been corroborated by a number of lab experiments (Andreoni and Vesterlund 2001; Eckel and Grossman 1998; Krawczyk 2010; Schildberg-Hörisch 2010; Kamas and Preston 2015) whilst others have found no difference (Bolton et al. 1998). My results suggest that the discrepancy with which men and women resolve equality-efficiency trade-offs increases with the salience of the moral context.

More broadly, I contribute to the growing economic literature on polarization. In the light of increasingly heated political debates over the extent and causes of global warming, the legitimacy of election results, and the perils of COVID-19, there has been increased interest in understanding the causes and consequences of partisan identity and polarization (Gentzkow et al. 2019; Canen et al. 2020). In behavioral economics, a prominent explanation for polarization has been the idea of motivated reasoning (Bénabou and Tirole 2016; Zimmermann 2020): When people are emotionally invested in a certain state of the world being true (e.g. because it favors them economically), it limits their ability to correctly update their beliefs. Extant research has documented the effect of motivated reasoning on ideological polarization Kahan (2013) and moral behavior (Grossman and Van Der Weele 2017). Fryer Jr et al. (2019) show that – depending on their prior beliefs – people interpret the same piece of scientific evidence drastically different. My findings suggest an additional behavioral mechanism: by drawing people's attention toward their divergent home-grown convictions of right and wrong, and away from selfish pragmatism as a common ground (Greene 2013), a political climate that emphasizes morality may contribute to polarization.

The next section introduces the theoretical framework and derives testable hypotheses. Section 3 explains the experimental design. Results are presented in section 4, and I discuss my findings in section 5.

2 Theoretical Framework

I assume that an individual i cares both about her material consumption opportunities $\pi(a_k)$ from choosing a certain action a_k, and about that action's compliance with her personal conception of morality $N_i(a_k, \widetilde{a}_i)$:[5]

$$u(a_k) = \phi \cdot V(\pi(a_k)) + (1-\phi) \cdot N_i(a_k, \widetilde{a}_i) \tag{1}$$

$V()$ denotes the utility from monetary payoffs and is increasing in $\pi(a_k)$. The function $N_i(a_k)$ compares the action a_k to one's personal moral values, and decreases as a_k deviates from i's conception of morally right behavior $\widetilde{a}_i$. The relative importance of the two utility sources is denoted by the weight $0 \leq \phi \leq 1$, capturing the moral wiggle room of the situation (Dana et al. 2007), i.e. the availability of plausible excuses for not adhering to $\widetilde{a}_i$ (Bénabou and Tirole 2011). While N_i is an idiosyncratic function capturing individual differences in people's inclination to comply with their respective normative ideals, ϕ is a situational parameter, which

[5] Similar approaches, assuming additively separable sources of utility are, for instance Brekke et al. (2003); Akerlof and Kranton (2000, 2005); Levitt and List (2007); Kessler and Milkman (2018); Burks and Krupka (2012); Krupka et al. (2017); see Vostroknutov (2020) for a recent review. Another strand of literature models the effect of (moral) identity on people's behavior as a Bayesian preference-signaling game (Bénabou and Tirole 2006; Bénabou and Tirole 2011; Grossman and van der Weele 2017; Kurschilgen and Marcin 2019).

can be externally influenced via the choice architecture. I denote a_i^* the action that maximizes material utility V, and a_i^{**} the action that maximizes overall utility u.

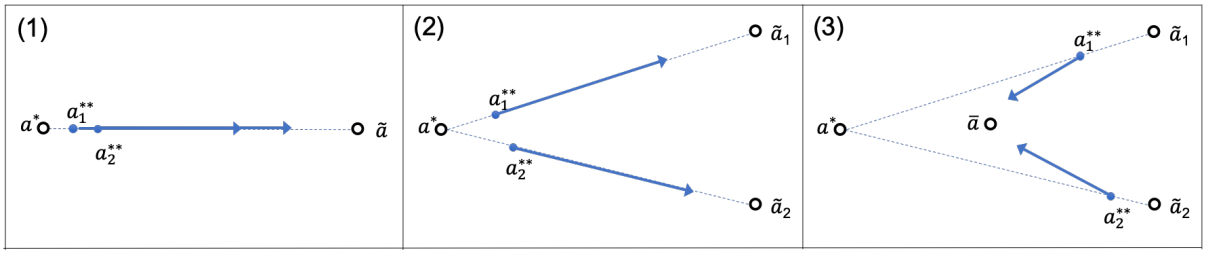

Figure 1: Behavioral Predictions

The effect of moral introspection is illustrated in Figure 1. When people agree on the moral ideal $\widetilde{a}$ (panel 1), differences in individual behavior are solely driven by people's home-grown attachment to morality, defined by the specifics of N_i. Introspection shrinks the moral wiggle room ϕ by increasing the salience of $\widetilde{a}$. As ϕ decreases, deviating from $\widetilde{a}$ becomes more expensive in utility terms, and consequently behavior becomes less selfish as it moves closer to $\widetilde{a}$. In contrast, when two individuals have different conceptions of the right thing to do, $\widetilde{a}_1 \neq \widetilde{a}_2$ (panel 2), introspection not only makes choices less selfish but also more polarized. As individuals' optimal choices move closer to their respective moral ideals, they move further away from one another, along the dimension of moral disagreement.

Polarization could be mitigated by people's desire for conforming with the behavior of their peers (Bernheim 1994; Michaeli and Spiro 2015). Assume they had information about the typical behavior of others, denoted by $\bar{a}$ in panel 3. In principle, information about others' behavior could affect a person's choice a_i^{**} both directly, as another additively-separable input of $u(a_k)$, or indirectly, by changing her perception of right and wrong, $\widetilde{a}_i$. If people's desire for conformity were sufficiently strong, in both cases, seeing how the average person – in a team, a role, an organization – behaves should have a homogenizing effect on people's behavior.

On the other hand, information about others' behavior could have the opposite impact if it mainly directed people's focus toward the discrepancy between what they observe and what they believe to be right (Cialdini et al. 1990; Krupka and Weber 2009), thereby reinforcing the weight of the latter. If the focusing effect were dominant, information disclosure would not mitigate polarization but could even exacerbate it.

3 Experimental Design

To test these predictions, I conduct a laboratory experiment with the following elements: (i) a choice environment in which people may plausibly pursue different moral ideals, (ii) a measure of people's underlying moral ideals, and (iii) an exogenous variation of moral introspection, and of social information.

Choice Environment. Participants play a modified dictator game (MDG) similar to Fisman et al. (2007) and Iriberri and Rey-Biel (2011, 2013). Each participant is assigned exactly one of two possible roles: one player is dictator i, the other is recipient j. Both players know their respective roles (i.e. no role-uncertainty) at the time the MDG is played. The recipient is passive. The dictator is confronted on her computer screen with four decision panels, each panel consisting of nine decision tasks (see Table 1), i.e. a total of 36 tasks. The four panels are shown sequentially to participants. Within each panel, the nine tasks are shown simultaneously. In every task, the dictator may choose between an option A and an option B. Option A is always profit maximizing. By choosing option B, at a cost of 10 tokens per task, a dictator can either – in panels 1 and 2 – create ($\pi_j^B > \pi_j^A$) additional income for the recipient or – in panels 3 and 4 – destroy ($\pi_j^B < \pi_j^A$) parts of it. In panels 1 and 3 the dictator is richer than the recipient ($\pi_i > \pi_j$), and in panels 2 and 4 she is poorer ($\pi_i < \pi_j$), independent of the specific option chosen.[6]

Table 1: Decision Panels

	(1) Ahead-Create		**(2) Behind-Create**		**(3) Ahead-Destroy**		**(4) Behind-Destroy**	
Task	π_i^A	π_j^A	π_i^A	π_j^A	π_i^A	π_j^A	π_i^A	π_j^A
1 to 9	170	70	110	120	140	130	90	180
Task	π_i^B	π_j^B	π_i^B	π_j^B	π_i^B	π_j^B	π_i^B	π_j^B
1	160	82	100	132	130	118	80	168
2	160	84	100	134	130	116	80	166
3	160	88	100	138	130	112	80	162
4	160	94	100	144	130	106	80	156
5	160	102	100	152	130	98	80	148
6	160	112	100	162	130	88	80	138
7	160	124	100	174	130	76	80	126
8	160	138	100	188	130	62	80	112
9	160	154	100	204	130	46	80	96

Note: Option A was constant across all 9 tasks of a given panel. For example, choosing B (instead of A) in task 3 of panel (1) meant paying 170-160=10 tokens in order to create 88-70=18 tokens for the other player.

If dictators are consistent, their choices in the MDG directly allow categorizing them in the two-dimensional Cartesian type space shown in Figure 2.[7] Perfectly selfish players are plotted

[6] The design of the MDG aims for subjects to make deliberate, well-thought choices in the spirit of the Holt and Laury (2002) test for risk attitudes. For that purpose, I deviate from the MDG of Iriberri and Rey-Biel (2011) in two respects: First, I let dictators choose between two options (Option A: selfish, Option B: destroy or create) instead of three (Option A: selfish, Option B: create, Option C: destroy). Second, instead of presenting the individual tasks randomly, I classify them into four panels and sort them within every panel by the relative price of creating/destroying.

[7] The term "consistent" refers to the General Axiom of Revealed Preferences (GARP). For further elaboration on GARP-consistency and on the type space, please see Appendix A.6.

in the origin, point S. The more a player deviates from purely selfish money maximizing, the greater the Cartesian distance to S. In particular, players may eye three prominent moral ideals:

(i) Social welfare, i.e. the maximization of total payoffs, irrespective of the particular distribution between dictator and recipient. To attain this ideal, a dictator would choose B in all tasks of panels 1 and 2 but choose A in all tasks of panels 3 and 4. In the Cartesian type space, this would be represented by point W.

(ii) Equality, i.e. the minimization of payoff differences between dictator and recipient. To attain this ideal, a dictator would choose B in all tasks of panels 1 and 4 but choose A in all tasks of panels 2 and 3, corresponding to point E.

(iii) Rawlsian Maximin, i.e. the maximization of the poorest player's payoffs (Rawls 1971). To attain this ideal, a dictator would choose B in all tasks of panel 1 but choose A in all tasks of panels 2, 3, and 4. In the type-space, such behavior would be depicted in point R.[8]

Measures. Through their decisions in the MDG, each player is characterized by one point in the type space of Figure 2. A person's opinion or choice is said to be more selfish (welfarists/egalitarian/Rawlsian) the closer her corresponding point in the type space is to the ideal point S (W/E/R).[9] To evaluate treatment differences, I will mainly focus on two aggregate measures: (A) Payoff sacrifice, measured as the mean deviation of choices from the selfish optimum (in tokens), i.e. the Cartesian distance between S and X in Figure 2; (B) Polarization, measured as the standard deviation of choices on the dimension of normative dissent. In the MDG, welfarist and egalitarian conceptions of morality are heavily at odds (and the Rawlsian ideal in between) when the dictator is *poorer* than the recipient. Consequently, normative polarization is captured by the *vertical* dispersion of choices in a given treatment.

[8] Other moral goals are also possible in the MDG. Since the aim of this paper is not to advance another test of distributional preferences but to examine how moral awareness affects behaviour in situations with divergent moral goals, I concentrate on the categories that have been found to be empirically most relevant in similar environments. For a very comprehensive and systematic overview of distributional archetypes see Kerschbamer (2015).

[9] Since the action space of the MDG is discrete, I interpret distance in the Cartesian sense. All results are robust to using Euclidean distance instead.

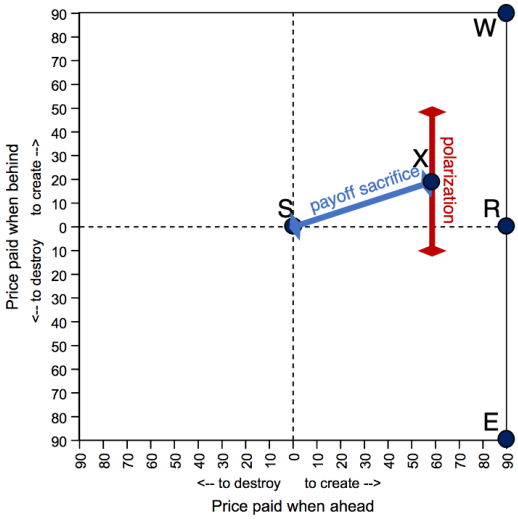

Figure 2: Type Space of the MDG

Note: Perfectly selfish players (i.e. who always choose option A in the MDG) are plotted in the origin, point S. The more a player deviates from payoff maximization, the larger the Cartesian distance to S. W denotes maximum social welfare, E denotes maximum equality, and R denotes compliance with Rawls' Maximin principle (Rawls 1971).

Treatments. Figure 3 provides an overview of the experimental setup. The *Baseline-u* treatment is the MDG described above. In treatment *Introspect-u*, subjects are prompted to state their moral opinion prior to making the incentivized choice. Specifically, after reading the experimental instructions but before being assigned the roles of dictator and recipient (i.e. behind the veil of ignorance), subjects are asked to privately state for each of the 36 tasks they will subsequently be seeing in the MDG: "Which of the two options (A or B) do you find morally right?" The instructions on the computer screen make it clear that the answers to this question are not payoff-relevant and will not be revealed to other participants.[10] After stating their moral opinions, subjects are assigned their roles and play the payoff-relevant MDG. When playing the MDG, dictators are reminded on their screens of their own, previously stated, moral opinions.

As a means to decrease subjects' moral wiggle room by raising their moral awareness, the moral opinion question is in a similar spirit as Krupka and Weber (2009), who have subjects deliberate about what others possibly said one should do ("injunctive focus"), as well as Gächter and Riedl (2005), who ask negotiators before a bargaining game to judge the situation from "the vantage point of a neutral arbitrator".[11]

To test for the effect of social information, I run two additional treatments, *Baseline-i* and *Introspect-i* (i="*informed*", u="*uninformed*"). In these treatments, participants are

[10] I deliberately do not incentivize the moral opinion questions. For a methodological discussion see Appendix A.7.

[11] In fact, the idea of moral introspection can be traced back to Adam Smith's, who called for strengthening one's moral self by becoming "the impartial spectator of one's own character and conduct" Smith (1790).

given salient and representative information about the behavior of participants in previous sessions. The information was taken from the real choice behavior of dictators in the *Baseline-u* treatment (see Appendix A.4).[12] In *Baseline-i*, participants are told that the experiment was run before with more than 100 participants, and that they will receive on their decision screens, information on how those previous participants chose in each of the 36 decision tasks of the MDG. Specifically, they see which of the two options (A or B) was chosen by the majority of previous dictators and how large the respective majority was (in%). This information was displayed on the decision screen next to each task. In *Introspect-i*, participants are shown the exact same information as in *Baseline-i* but they receive it already at the Moral Opinion stage, and then again at the Choice stage.

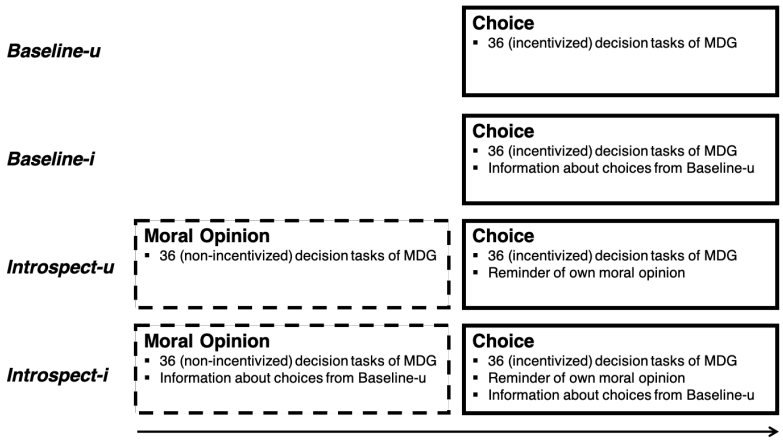

Figure 3: Experimental Setup

Procedures. The experiment was conducted at the BonnEconLab, Germany. Subjects were recruited via email from a pool of more than 5000 people, using the software ORSEE (Greiner 2015). In a between-subject design, 640 participants (320 dictators) took part in the experiment. Participants were mainly University of Bonn undergraduates from a variety of disciplines, 58% were female. Participants were seated in visually completely isolated cubicles. Paper instructions (see Appendix A.1) explaining the MDG were identical in all treatments. They were handed out to the participants while they were sitting in their cubicles, and were subsequently read aloud by the experimenter.

At the end of experiment, the computer randomly picked one decision task per panel for payoff. The corresponding token amounts from those four decision tasks were added and converted into Euros (100 tokens = € 1). Participants earned on average € 6 (ca. US$ 8) for approximately 20 minutes of lab-time, which corresponds to about twice the typical student's hourly wage. Immediately after the MDG, subjects answered a non-incentivized questionnaire, covering socio-demographics (age, gender, number of siblings), stated risk and trust attitudes as commonly elicited in the German Socioeconomic Panel (SOEP), and the Big-Five person-

[12] For a similar implementation of social information, see Engel et al. (2021)

ality traits (extraversion, agreeableness, conscientiousness, neuroticism, openness) according to Rammstedt and John (2007). The experiment was computerized in ztree (Fischbacher 2007). Table 2 summarizes the data collected in the experiment.[13]

Table 2: Data Structure of the Experiment

	sessions	subjects	deciders	GARP-consistent
Baseline-u	13	304	152	120
Baseline-i	6	144	72	63
Introspect-u	4	96	48	39
Introspect-i	4	96	48	44
	27	640	320	266

Note: The disproportionately large number of participants in *Baseline-u* is due to the need for collecting a large number of choices in order to provide sufficiently representative information to subjects in *Baseline-i* and *Introspect-i*.

4 Results

I first report causal evidence on how moral introspection affects behavior in the MDG, and show that the effect is robust to subjects having information about the descriptive norm. Subsequently, I examine how individual moral opinions translate into actual incentivised behavior, and show that introspection has a systematically different effect on men than on women, exacerbating existing normative discrepancies between genders.

Introspection, information, and behavior. To test how moral introspection affects behavior in the MDG, I compare deciders' individual choices using the two main measures defined in the preceding section: payoff sacrifice and polarization. The results are summarized in Table 3. Panel A shows that when deciders were not informed about typical behavior in the MDG, and were not asked to state their moral opinion prior to playing the incentivized MDG (*Baseline-u*), they sacrificed on average 37 tokens of monetary payoff. Asking them to state their moral opinion under the veil of ignorance (*Introspect-u*) almost doubles the monetary payoff they are willing to sacrifice (p=.005). This result corroborates extant research showing that moral nudges make people less selfish, and extends it to normatively ambivalent situations. More

[13] I only consider the results of dictators whose choice behavior in the MDG was consistent with GARP. For the uninformed players, the percentage of consistent dictators is virtually identical (Probit regression, p=.73) in *Baseline-u* (79%) and *Introspect-u* (81%). Also for the informed players, the share of consistent dictators does not vary significantly (Probit regression, p=.472) across treatments: 88 in *Baseline-i* and 92 in *Introspect-i*. Including inconsistent dictators would require additional assumptions about the interpretation of dictators' "errors". For further explanation of GARP-consistency in the MDG, see Appendix A.6. For a similar procedure, see for instance Sutter et al. (2013).

interestingly, however, Panel B shows that moral introspection also increases the polarization of incentivized choices by 33% (p=.002).

Table 3: Treatment Effects on Choices in MDG

(A) Payoff Sacrifice

	uninformed	*informed*	
Baseline	37	32	p=.433
Introspect	71	60	p=.555
	p=.005	p=.006	

(B) Polarization

	uninformed	*informed*	
Baseline	33	28	p=.930
Introspect	44	37	p=.236
	p=.002	p=.027	

(A) Payoff Sacrifice is measured as the mean deviation of choices from the selfish optimum (in tokens). P-values from two-sided Mann-Whitney ranksum tests. (B) Polarization is measured as the standard deviation of choices when the decider is behind (the *vertical* dimension in Figure 2). When the decider is behind, she could sacrifice payoff for increasing total surplus while accepting higher inequality, or for increasing equality while accepting lower total surplus. P-values from two-sided Levene tests.

If people's desire for conformity were pronounced, information about others' behavior could mitigate polarization. I test the effect of unbiased information disclosure in treatments *Baseline-i* and *Introspect-i*, in which participants are given representative information on how others decided in each of the 36 decision tasks of the MDG (see Table A1 for details). Since the information was obtained from the true behavior of participants in *Baseline-u*, it showed behavior that was predominantly selfish. In every single task, participants saw that a majority (between 53% and 99%) had chosen the payoff maximizing option A. They also saw that – when deciders were richer than the recipient – 47% were willing to sacrifice some payoffs to create additional income for the recipient (consistent with all three normative ideals). When deciders in *Baseline-u* were poorer than the recipient, they were rather inclined toward sacrificing money to increase welfare (33%) than to increase equality (11%).

My findings do not reveal any conformity effect. Despite seeing high levels of selfishness, participants in *Introspect-i* were not significantly more selfish than in *Introspect-u* (p=.555). Similarly, the information disclosure was not capable of significantly reducing polarization (p=.236). In contrast, I find a similarly strong effect of moral introspection for the informed participants as for the uninformed ones. The willingness to sacrifice payoffs increases by 86% (p=.006) and polarization by 32% (p=.027).

Moral opinions and subsequent behavior. As with incentivized choices, social information has no effect on participants' opinions of morally right behavior. Moral opinions in *Introspect-i* are neither more selfish (Mann-Whitney ranksum test (MW), two-sided, N=83, p=.171), nor more welfare oriented (p=.219) than in *Introspect-u*. There is also no difference with respect to their proximity to the Rawlsian ideal (p=.541), nor to the egalitarian ideal (p=.312). Since being

informed about the descriptive norm of the situation is manifestly irrelevant in this setting, for the subsequent analysis I pool the data from *Baseline-i* and *Baseline-u* (and simply refer to it as *Baseline*), as well as the data from *Introspect-i* and *Introspect-u* (into *Introspect*).[14]

The right panel of Figure 4 depicts the moral opinions in the *Introspect* treatments. The size of the bubbles indicates the percentage of (subsequent) dictators located in a given point.[15] The four largest bubbles correspond exactly to the ideal points described in Figure 4: pure selfish (14%), pure welfarist (16%), pure Rawlsian (8%) and pure egalitarian (6%). In total, 47% of dictators are on the right boundary of the graph. Note the extent of dissimilarity between the moral opinions of these players, who are furthest away from the selfish optimum. The welfare-oriented players depicted in the upper-right corner believe it is "morally right" to prefer allocation (100, 204) over (110, 120), i.e. to create 74 units of additional wealth even if this means increasing inequality by 94 units. In sharp contrast, the egalitarian players depicted in the lower-right corner believe it is "morally right" to prefer allocation (80, 96) over (90, 180), i.e. to reduce inequality by 74 units even if this means destroying wealth by 94 units.

How do moral opinions translate into actual behavior once the veil of ignorance is lifted? The model predicts them to become more selfish (i.e. reduce the distance to the origin) but to not change their general moral orientation (i.e. the angle). To answer this question, I first classify dictators into different types according to their moral opinions, and then analyse – type by type – the individual transition from opinion (right panel of Figure 4) to choice (center panel) within the *Introspect* treatment. 47% of dictators classify as welfarists as they judge it morally right to create income for the recipient both when being richer and poorer (i.e. the upper-right quadrant), 18% as Rawlsians (i.e. the horizontal line between the two right-hand quadrants), 18% as egalitarians (i.e. the lower-right quadrant), and 14% as selfish (i.e. the origin).

[14] For a similar procedure, see for instance Hamman et al. (2010). All reported results also hold if I do not pool the data but rather analyze separately for informed and uniformed dictators. But pooling increases the statistical power.

[15] Remember that at this point, the players did not know whether they would be a dictator or a recipient in the subsequent MDG.

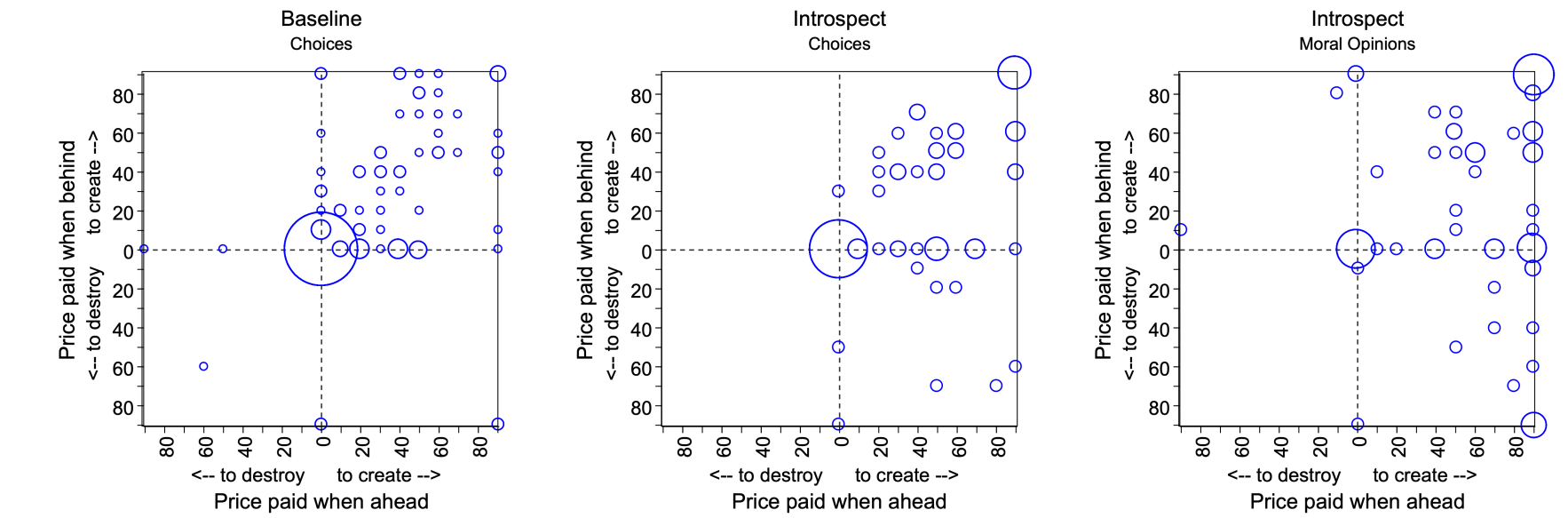

Figure 4: Individual Choices and Opinions

Note: The left panel shows the choices in *Baseline*, the center panel the choices in *Introspect* and the right panel the moral opinions in *Introspect*. The area of the circles is proportional to the relative frequency of observations at a given point in the type space.

Faced with the incentivized choice, the welfarist dictators become significantly more selfish; the mean distance from point S (as defined in Figure 2) decreases from 137 to 96 tokens (Wilcoxon signrank test (WSR), two-sided, N=39, p<.001) just as the distance from W increases from 43 to 84 tokens (p<.001). But these dictators do not revise their general moral orientation as their distance from R (p=.241) and E (p=.241) stays virtually unchanged. Similarly, the egalitarian dictators display significantly more selfish – from 129 to 66 tokens (WSR, two-sided, N=15, p=.002) – and less egalitarian choices – from 51 to 122 tokens (p=.002). But they do not become more or less welfare oriented (p=.871) nor Rawlsian (p=.262). The Rawlsian dictators display a similar pattern as their choices move closer to S (WSR, two-sided, N=15, p=.013) and away from R (p=.001), but without altering their vertical orientation (p=.157). Finally, all but one of the selfish dictators confirm their moral opinion with an identical subsequent choice.

Comparing the right panel and the center panel of Figure 4, it is striking how the rather extended type space of moral opinions translates into a much narrower type space of actual choices. There is thus a substantial moral opinion-moral action gap (Rest 1986; Ellertson et al. 2016). On average, moving from moral opinion to incentivized choice in *Introspect* decreases the distance from S by 37 tokens, (WSR, two-sided, N=83, p<.001) while increasing the distance from W by 23 tokens (p<.001), from R by 7 tokens (p=.011), and from E by 19 tokens (p<.001). Overall, moral opinions are significantly more polarized than subsequent choices, as the vertical dispersion decreases from 52 to 40 tokens (Levene test, two-sided, N=83, p=.009). This supports the fundamental assumption of the model presented above that people have, independent of their individual conception of the moral ideal, selfishness as a common denominator. As dictators trade off compliance with their individual moral ideal against selfish profit maximization, their actual incentivized choices become less polarized than the moral opinions they stated behind the veil of ignorance. And yet, as we saw in Table 3, their choices after moral introspection are significantly less selfish and more polarized than in the *Baseline*.

Gender Differences. We conjectured that raising moral introspection would increase polarization by exacerbating existing normative discrepancies. As we have seen in the above analysis, there are indeed substantial, home-grown discrepancies in participant's conceptions of right and wrong. Those existing normative discrepancies impact behavior more strongly when people's moral awareness is raised. In social groups (teams, organizations, societies), there may be many reasons for people having different home-grown conceptions of right and wrong. In addition to differences of gender, a social group might span several generations, people could have different cultural or religious heritage, and various educational backgrounds. In contrast, the participants of this experiment are rather homogeneous. They are predominantly German university students, aged 20 to 29. There is, however, a rather even gender split. Several studies show that women are more inclined towards equality whereas men favor social welfare as normative criterion (Andreoni and Vesterlund 2001; Eckel and Grossman 1998; Krawczyk 2010; Schildberg-

Hörisch 2010; Kamas and Preston 2015) whilst others have found no difference (Bolton et al. 1998).

Table 4: Gender Effects

	S-distance		**W**-distance		**R**-distance		**E**-distance	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
Introspect	31.08***	47.39***	-28.09***	-46.27***	-8.059**	1.253	-14.09***	-0.466
	(7.846)	(12.46)	(8.131)	(13.33)	(3.612)	(5.199)	(4.606)	(5.852)
Female	-13.04*	-3.894	14.48**	4.293	-1.960	3.258	-0.411	7.222
	(7.186)	(7.884)	(7.335)	(8.112)	(3.487)	(4.332)	(4.583)	(5.385)
Introspect × *Female*		-29.45*		32.82**		-16.81**		-24.59***
		(15.48)		(16.20)		(7.346)		(9.085)
Informed	-5.432	-5.121	-1.693	-2.040	-0.247	-0.0698	5.803	6.063
	(6.849)	(6.773)	(7.053)	(6.954)	(3.257)	(3.214)	(3.898)	(3.871)
Age	0.618	0.679	-0.582	-0.650	0.0677	0.103	-0.210	-0.159
	(0.811)	(0.791)	(0.795)	(0.783)	(0.419)	(0.437)	(0.569)	(0.598)
Siblings	0.190	0.626	-1.926	-2.412	-0.613	-0.364	-0.752	-0.388
	(3.663)	(3.653)	(3.765)	(3.690)	(1.549)	(1.548)	(2.384)	(2.321)
Risk	-0.0469	0.186	1.403	1.144	-0.454	-0.321	-1.166	-0.971
	(1.482)	(1.487)	(1.588)	(1.618)	(1.015)	(1.029)	(1.046)	(1.045)
Trust	4.796**	5.184**	-3.960*	-4.392**	0.119	0.341	-0.352	-0.0276
	(2.105)	(2.105)	(2.156)	(2.156)	(0.932)	(0.938)	(1.161)	(1.139)
Extraversion	-5.921***	-6.101***	5.237***	5.437***	-0.0976	-0.200	-0.293	-0.443
	(1.917)	(1.897)	(1.955)	(1.943)	(1.036)	(1.044)	(1.194)	(1.199)
Agreeableness	3.691*	3.878*	-4.869**	-5.077**	0.0987	0.205	0.949	1.105
	(2.185)	(2.178)	(2.382)	(2.334)	(1.259)	(1.256)	(1.393)	(1.379)
Conscientiousness	0.106	0.287	-0.593	-0.795	-0.0772	0.0264	1.119	1.271
	(1.992)	(1.983)	(1.996)	(1.971)	(0.826)	(0.837)	(1.117)	(1.122)
Neuroticism	-1.502	-1.574	2.148	2.228	-0.313	-0.354	-1.313	-1.373
	(1.663)	(1.690)	(1.704)	(1.729)	(0.966)	(0.950)	(1.293)	(1.277)
Openness	3.383*	3.325*	-0.990	-0.925	0.829	0.796	-0.708	-0.757
	(1.854)	(1.860)	(1.858)	(1.870)	(1.022)	(0.986)	(1.248)	(1.215)
Constant	-4.139	-15.45	175.1***	187.7***	88.64***	82.18***	190.3***	180.8***
	(35.52)	(35.14)	(35.09)	(34.79)	(17.13)	(17.26)	(20.91)	(20.52)
Observations	266	266	266	266	266	266	266	266
R-squared	0.164	0.178	0.138	0.155	0.026	0.046	0.057	0.084

Ordinary Least Squares regression. The dependent variables in each set of columns denote the Cartesian distance from the ideal points S (selfish), W (welfare), R (Rawls), and E (equality). Robust standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4 shows the results from an OLS estimation. Confirming the results reported above, Column (1a) shows that after controlling for socio-demographics, risk and trust attitudes, as well as Big-Five personality traits, choices in *Introspect* are 31.08 tokens further away from S than in *Baseline*, i.e. less selfish. Columns (2a), (3a), and (4a) show that in turn choices decrease the Cartesian distance towards the moral ideals of welfare (by 28.09 tokens), Rawlsian Maximin (8.06 tokens), and equality (14.09 tokens). More interestingly, the interaction *Introspect* × *Female* reveals that raising moral awareness has a pronounced gender compo-

nent. In fact, the general increase in polarization observed in *Introspect* coincides largely with an increase of gender differences. Column (2b) shows that whilst moral introspection makes men substantially more welfare-oriented (by 46.27 tokens), this effect is significantly smaller for women (32.82 tokens less) and, in sum, insignificantly different from zero (p=.156). In sharp contrast, columns (2b) and (3b) indicate that moral introspection makes women's choices significantly more Rawlsian (by 15.56 tokens) and equality-oriented (25.06 tokens) whereas the effect for men is virtually zero.

5 Discussion

I report experimental evidence showing that, if people have home-grown discrepancies about which type of behavior is (morally) right and wrong, moral introspection makes choices not only less selfish but also more polarized, as people stick closer to their respective moral *extremes*. Providing people with social information about the behavior of their peers is not capable of homogenizing behavior. In the – normatively ambivalent – allocation setting studied in this paper, I find that increasing moral awareness exacerbates gender differences as men value welfare stronger whilst women put more weight on equality.

Allocation problems are not only of obvious importance to many applications, ranging from wage negotiations and distributions of bonuses, to questions of fair taxation, to issues of political representation, they also have the interesting property to feature more than one plausible moral ideal. Future research should extend the present analysis to other settings in which people disagree on the right thing to do (merit vs. necessity, freedom vs. security, pro-life vs. pro-choice, etc.), and to other subject populations.

A growing strand of research in the social sciences is concerned with finding ways to restrain people's selfishness and encourage their moral responsibility. Recent work has suggested that raising moral awareness – and thus narrowing moral wiggle room – might be an avenue of improving social outcomes. My findings (i) add new evidence that raising awareness reduces selfish behavior, and (ii) point at a potentially important caveat: When people differ with respect to their particular moral ideals, as participants do already in the simple, non-strategic, laboratory environment of the present study, appealing to morality might actually increase the difficulty of finding consensual solutions to social problems. My findings suggest that there might sometimes be a social cost to moral sensitivity. Further research should test this conjecture in situations of strategic interaction, particularly in problems of coordinating conflicting interests, like wage negotiations. When people disagree on moral goals, they might actually find some common ground in human selfishness. In that sense, future work should attempt to identify whether in certain situations appealing to material selfishness may, ironically, actually lead to better social outcomes.

More broadly, my results demonstrate that an exogenous variation of people's awareness

about a situation's moral implications can have important effects on people's behavior. Both within organizations and in the political arena, however, the moral frame of an issue (and even the general level of moral sensitivity), is typically an endogenous variable, i.e. a variable that can be influenced by strategic players (politicians, activists, managers, etc.). Future work should study how moral awareness is used (and possibly abused) as a strategic variable.

The results of this paper should, of course, not be taken as arguments against moral nudges in general. Rather, my findings suggest that there is an additional layer of difficulty choice architects need to consider when trying to develop a healthy moral culture within a social group.

References

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

Akerlof, G. A. and R. E. Kranton (2005). Identity and the economics of organizations. *Journal of Economic Perspectives 19*(1), 9–32.

Alesina, A. and E. L. Ferrara (2005). Preferences for redistribution in the land of opportunities. *Journal of Public Economics 89*(5), 897 – 931.

Andreoni, J. and L. Vesterlund (2001). Which is the fair sex? gender differences in altruism. *The Quarterly Journal of Economics 116*(1), 293–312.

Bartling, B. and U. Fischbacher (2011). Shifting the blame: On delegation and responsibility. *The Review of Economic Studies 79*(1), 67–87.

Bartling, B., R. A. Weber, and L. Yao (2014). Do markets erode social responsibility? *The Quarterly Journal of Economics 130*(1), 219–266.

Bašić, Z. and E. Verrina (2020). Personal norms—and not only social norms—shape economic behavior. *MPI Collective Goods Discussion Paper* (2020/25).

Bénabou, R., A. Falk, and J. Tirole (2019). Narratives, imperatives, and moral persuasion. *NBER working paper*.

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

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

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

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy 102*(5), 841–877.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms.* Oxford University Press.

Bicchieri, C. and E. Dimant (2019). Nudging with care: The risks and benefits of social information. *Public choice*, 1–22.

Bicchieri, C. and E. Xiao (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making 22*(2), 191–208.

Bolton, G. E., E. Katok, and R. Zwick (1998). Dictator game giving: Rules of fairness versus acts of kindness. *International Journal of Game Theory 27*(2), 269–299.

Brekke, K. A., S. Kverndokk, and K. Nyborg (2003). An economic model of moral motivation. *Journal of Public Economics 87*(9-10), 1967–1983.

Burks, S. V. and E. L. Krupka (2012). A multimethod approach to identifying norms and normative expectations within a corporate hierarchy: Evidence from the financial services industry. *Management Science 58*(1), 203–217.

Canen, N., C. Kendall, and F. Trebbi (2020). Unbundling polarization. *Econometrica 88*(3), 1197–1233.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics 117*(3), 817–869.

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

Coca-Cola (2018). Code of business conduct. https://www.coca-colacompany.com/content/dam/journey/us/en/private/fileassets/pdf/2018/Coca-Cola-COC-External.pdf.

Corneo, G. and H. P. Grüner (2002). Individual preferences for political redistribution. *Journal of Public Economics 83*(1), 83–107.

Dal Bó, E. and P. Dal Bó (2014). "Do the right thing:" the effects of moral suasion on cooperation. *Journal of Public Economics 117*, 28–38.

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

Di Tella, R., R. Perez-Truglia, A. Babino, and M. Sigman (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *American Economic Review 105*(11), 3416–42.

Dimant, E. (2021). Hate trumps love: The impact of political polarization on social preferences. *mimeo*.

Eckel, C. C. and P. J. Grossman (1998). Are women less selfish than men?: Evidence from dictator experiments. *The Economic Journal 108*(448), 726–735.

Ellertson, C. F., M.-C. Ingerson, and R. N. Williams (2016). Behavioral ethics: A critique and a proposal. *Journal of Business Ethics 138*(1), 145–159.

Engel, C., S. Kube, and M. Kurschilgen (2021). Managing expectations: How selective information affects cooperation and punishment in social dilemma games. *Journal of Economic Behavior & Organization 187*, 111–136.

Engel, C. and M. Kurschilgen (2020). The fragility of a nudge: the power of self-set norms to contain a social dilemma. *Journal of Economic Psychology 81*, 102293.

Falk, A. and N. Szech (2013). Morals and markets. *Science 340*(6133), 707–711.

Feldman, Y. and E. Halali (2017). Regulating "good" people in subtle conflicts of interest situations. *Journal of Business Ethics*.

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

Fisman, R., S. Kariv, and D. Markovits (2007). Individual preferences for giving. *American Economic Review 97*(5), 1858–1876.

Fong, C. (2001). Social preferences, self-interest, and the demand for redistribution. *Journal of Public Economics 82*(2), 225–246.

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

Friesdorf, R., P. Conway, and B. Gawronski (2015). Gender differences in responses to moral dilemmas. *Personality and Social Psychology Bulletin 41*(5), 696–713.

Fryer Jr, R. G., P. Harms, and M. O. Jackson (2019). Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association 17*(5), 1470–1501.

Gächter, S. and A. Riedl (2005). Moral property rights in bargaining with infeasible claims. *Management Science 51*(2), 249–263.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica 87*(4), 1307–1340.

Gino, F., M. E. Schweitzer, N. L. Mead, and D. Ariely (2011). Unable to resist temptation: How self-control depletion promotes unethical behavior. *Organizational Behavior and Human Decision Processes 115*(2), 191–203.

Google (2018). Code of conduct. https://abc.xyz/investor/other/google-code-of-conduct.html.

Graham, J., J. Haidt, and B. A. Nosek (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology 96*(5), 1029–1046.

Graham, J., P. Meindl, E. Beall, K. M. Johnson, and L. Zhang (2016). Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology 8*, 125–130.

Graham, M. H. and M. W. Svolik (2020). Democracy in america? partisanship, polarization, and the robustness of support for democracy in the united states. *American Political Science Review 114*(2), 392–409.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

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

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

Grossman, Z. and J. J. van der Weele (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association 15*(1), 173–217.

Haidt, J. and J. Graham (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research 20*(1), 98–116.

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

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*(5), 1644–1655.

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

Iriberri, N. and P. Rey-Biel (2013). Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do? *Quantitative Economics 4*(3), 515–547.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision Making 8*, 407–24.

Kamas, L. and A. Preston (2015). Can social preferences explain gender differences in economic behavior? *Journal of Economic Behavior & Organization 116*, 525–539.

Kaptein, M. (2015). The effectiveness of ethics programs: The role of scope, composition, and sequence. *Journal of Business Ethics 132*(2), 415–431.

Kerschbamer, R. (2015). The geometry of distributional preferences and a non-parametric identification approach: The equality equivalence test. *European Economic Review 76*, 85–103.

Kessler, J. B. and K. L. Milkman (2018). Identity in charitable giving. *Management Science 64*(2), 845–859.

Kouchaki, M. and I. H. Smith (2014). The morning morality effect: The influence of time of day on unethical behavior. *Psychological Science 25*(1), 95–102.

Krawczyk, M. (2010). A glimpse through the veil of ignorance: Equality of opportunity and support for redistribution. *Journal of Public Economics 94*(1-2), 131–141.

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

Krupka, E. L., S. Leider, and M. Jiang (2017). A meeting of the minds: Informal agreements and social norms. *Management Science 63*(6), 1708–1729.

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

Kurschilgen, M. and I. Marcin (2019). Communication is more than information sharing: The role of status-relevant knowledge. *Games and Economic Behavior 113*, 651–672.

Levitt, S. D. and J. A. List (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives 21*(2), 153–174.

Lindström, B., S. Jangard, I. Selbing, and A. Olsson (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General 147*(2), 228.

Michaeli, M. and D. Spiro (2015). Norm conformity across societies. *Journal of Public Economics 132*, 51–65.

Müller, D. and S. Renes (2021). Fairness views and political preferences: evidence from a large and heterogeneous sample. *Social Choice and Welfare 56*(4), 679–711.

Rammstedt, B. and O. P. John (2007). 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 41*(1), 203–212.

Rapp, C. (2016). Moral opinion polarization and the erosion of trust. *Social Science Fesearch 58*, 34–45.

Ravallion, M. and M. Lokshin (2000). Who wants to redistribute?: The tunnel effect in 1990s russia. *Journal of Public Economics 76*(1), 87–104.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Rest, J. R. (1986). Moral development: Advances in research and theory.

Schildberg-Hörisch, H. (2010). Is the veil of ignorance only a concept about risk? an experiment. *Journal of Public Economics 94*(11-12), 1062–1066.

Schram, A. and G. Charness (2015). Inducing social norms in laboratory allocation choices. *Management Science 61*(7), 1531–1546.

Shang, J. and R. Croson (2009). A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal 119*(540), 1422–1439.

Smith, A. (1790). *The Theory of Moral Sentiments: Or, An Essay Towards an Analysis of the Principles by which Men Naturally Judge... To which is Added a Dissertation on the Origin of Languages*, Volume 1. Strahan.

Sutter, M., M. G. Kocher, D. Glätzle-Rüetzler, and S. T. Trautmann (2013). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review 103*(1), 510–31.

Svolik, M. W. et al. (2020). When polarization trumps civic virtue: Partisan conflict and the subversion of democracy by incumbents. *Quarterly Journal of Political Science 15*(1), 3–31.

te Velde, V. L. (2020). Heterogeneous norms: Social image and social pressure when people disagree.

Vostroknutov, A. (2020). Social norms in experimental economics: Towards a unified theory of normative decision making. *Analyse & Kritik 42*(1), 3–39.

Welsh, D. T. and L. D. Ordóñez (2014). Conscience without cognition: The effects of subconscious priming on ethical behavior. *Academy of Management Journal 57*(3), 723–742.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review 110*(2), 337–61.

A Online-Appendix

A.1 Paper instructions

General Information

Welcome to our experiment!
If you read the following explanations carefully, you will be able to earn a substantial sum of money, depending on the decisions you make. It is therefore crucial that you read these explanations carefully.
During the experiment there shall be absolutely no communication between participants. Any violation of this rule means you will be excluded from the experiment and from any payments. If you have any questions, please raise your hand. We will then come over to you.
During the experiment we will not calculate in euro, but instead in tokens. Your total income is therefore initially calculated in tokens. The total number of tokens you accumulate in the course of the experiment will be transferred into Euro at the end, at a rate of

100 tokens = 1 Euro

At the end you will receive from us the **cash** sum, in euro, based on the number of tokens you have earned.

The Experiment

In the experiment, there are two roles: **decider** and **recipient**.
At the beginning of the experiment you will be randomly allotted one of the two roles. One half of the participants will be deciders, the other half will be recipients. During the entire experiment, you will remain in the same role.
On your computer screen you will be shown **4 tables**, one after the other. Every table consists of **9 decision tasks**.
A decision task could for example read as follows:

	Option A		Option B	
Decider (You)	12		10	
Recipient	5		7	Your decision (A or B):

In every decision task the decider has to choose between **Option A** and **Option B**. The two options define how many **tokens** the decider gets and how many the recipient gets.
In this example the decider gets 12 tokens and the recipient 5 tokens if the decider chooses Option A. If the decider chooses Option B, the decider gets 10 tokens and the recipient 7 tokens.
In every decision task the computer will **randomly** match every decider with a different recipient. Thus, the decider-recipient pairs change in every decision task.
The decider will never know the identity of the recipient.
The recipient will never know the identity of the decider.
At the end of every table please press the "OK" button on the lower right hand side of your screen. Only after pressing "OK" your decisions are saved and become effective. You will then be shown the next table.

Payoffs

At the end of experiment the computer will **randomly** pick one decision task out of every table. The computer thus picks in total **4 decision tasks**, one from every table. The corresponding token amounts from those 4 decision tasks will be added and changed into Euros.
If you are **decider**, your payoffs only depend on your own choices and on the random draw at the end of the experiment.
If you are **recipient**, your payoffs only depend on the choices of the corresponding decider and the random draw at the end of the experiment.

A.2 Additional screen in *Introspect* treatments

Before the computer randomly determines who will be Decider and who will be Recipient, we would like to know your opinion.
We would like to know from you:
Which of the two Options (A or B) do you find morally right?
The answers to these questions will be kept anonymous. No other participant will get to know them at any time.
Your answers to these questions are not relevant for your payoffs.

A.3 Additional screen in the *informed* treatments

This Experiment has been run before with more than 100 Deciders.
In the column on the right-hand side of your screen you can see how the Deciders in those previous Experiments decided. Specifically, you will be shown which percentage of Deciders chose Option A or Option B in the corresponding Choice Task.

A.4 Information about choices of previous players

Table A1 displays the information given to players in the *Baseline-i* and *Introspect-i* treatments. The information represents the percentage of deciders choosing the majority option in the *Baseline-u* treatment.

Table A1: Information given in the *informed* treatments

Panel Task	Ahead – Create	Ahead – Destroy	Behind – Create	Behind – Destroy
1	89% chose A	95% chose A	91% chose A	95% chose A
2	89% chose A	95% chose A	92% chose A	91% chose A
3	87% chose A	97% chose A	89% chose A	95% chose A
4	83% chose A	98% chose A	88% chose A	90% chose A
5	76% chose A	99% chose A	84% chose A	90% chose A
6	68% chose A	97% chose A	76% chose A	89% chose A
7	64% chose A	97% chose A	74% chose A	91% chose A
8	58% chose A	97% chose A	68% chose A	90% chose A
9	53% chose A	97% chose A	67% chose A	89% chose A

A.5 Post-experimental tests

Risk. The risk question reads: "Are you, generally speaking, a person willing to take risks or do you rather try to avoid risks?" (0-not at all willing to take risks, ... , 10-very willing to take risks).

Trust. The trust questions read: "Please rate the following three statements on a scale from 1 to 4 (1-fully agree, 2-rather agree, 3-rather disagree, 4-fully disagree): (A) Generally, people can be trusted. (B) Nowadays you cannot trust anybody. (C) When dealing with strangers it's better to be careful before trusting them." The composite trust measure is (5-A)+B+C and ranges from 3 (low trust) to 12 (high trust).

A.6 Modified Dictator Game (MDG)

Subjects choices in the MDG allow fitting the parameters of the Charness and Rabin (2002) model of other-regarding preferences:

$$U_i(a_i) = \begin{cases} (1-\rho)\pi_i(a_i) + \rho\pi_j(a_i) & \text{if } \pi_i \geq \pi_j \\ (1-\sigma)\pi_i(a_i) + \sigma\pi_j(a_i) & \text{if } \pi_i \leq \pi_j \end{cases} \tag{1}$$

An individual i's utility $U_i(\cdot)$ from taking a certain action a_i depends on the action's payoff consequences for herself and for another player j. Every person is characterized by her concern for others when she is richer (ρ) and when she is poorer (σ). Player i prefers allocation B to

allocation A if $U_i^B \geq U_i^A$. Assuming Charness and Rabin (2002) preferences and $\pi_i \geq \pi_j$ this implies:

$$\pi_i^B - \pi_i^A \geq \rho((\pi_i^B - \pi_i^A) - (\pi_j^B - \pi_j^A)) \tag{A2}$$

For convenience, I define $\Delta_i = \pi_i^B - \pi_i^A$ to obtain:

$$\Delta_i \geq \rho(\Delta_i - \Delta_j) \tag{A3}$$

The same argument applies to $\pi_i \leq \pi_j$ by simply replacing ρ with σ. Assuming $\Delta_i < 0$ (i.e. allocation B is less profitable to player i than allocation A) a person's choice reveals her ρ and σ parameters as depicted in Table A2.

Table A2: Parameter Space of the MDG

		Ahead	Behind
		$\pi_i \geq \pi_j$	$\pi_i \leq \pi_j$
Create	$\Delta_i < \Delta_j$	$\rho \geq \frac{\Delta_i}{\Delta_i - \Delta_j} > 0$	$\sigma \geq \frac{\Delta_i}{\Delta_i - \Delta_j} > 0$
Destroy	$\Delta_i > \Delta_j$	$\rho \leq \frac{\Delta_i}{\Delta_i - \Delta_j} < 0$	$\sigma \leq \frac{\Delta_i}{\Delta_i - \Delta_j} < 0$

Table A3 illustrates how the experimental MDG devotes one decision panel to each of these four situations. In the two panels on the left, the dictator's payoff is always higher than the recipient's ($\pi_i \geq \pi_j$) whereas in the two right-hand panels it is the other way around ($\pi_i \leq \pi_j$). In the two upper panels the dictator can create income for the recipient ($\Delta_i < 0 < \Delta_j$) whereas in the two lower panels she can reduce the recipient's income ($0 > \Delta_i > \Delta_j$).

In each panel k, there are nine decision tasks t. In each task the dictator has to choose between option A and option B, specifying two different payoff allocations for the dictator and the corresponding recipient. Option A is the same for every task within a given panel. Option B creates or destroys income of the recipient at a cost of 10 tokens. Take for example task 1 of the Ahead-Create panel. If the dictator chooses option A she receives 170 tokens and the recipient 70 tokens, and if she chooses option B she gets 160 and the recipient 82.

In each panel the relative price of creating/destroying decreases with every task. In task 1, the dictator has to give up 10 tokens to create/destroy 12 tokens whereas in task 9 for the same cost the dictator creates/destroys 84 tokens. Consequently, choosing option B in task 1 and option A in task 2 of the same panel would violate the General Axiom of Revealed Preferences (GARP). In the MDG, a GARP-consistent dictator should have at most one switch from Option

A to option B per panel, and no switch from B to A. In addition, consistency requires players not to both create and destroy when they are ahead (or behind). If these consistency requirements are met, the parameters ρ and σ of a given dictator are defined by the point at which she switches from option A to option B.

In Figure 2, dictators who always choose the profit maximizing Option A have revealed having both Charness and Rabin (2002) parameters close to zero, more specifically $-.14 \leq \rho \leq .11$, and $-.14 \leq \sigma \leq .11$. Consequently, those players are located at the origin of the graph, denoted by point S. Every time a dictator chooses option B over option A she pays a price of 10 tokens. According to GARP, there should only be at most one switch from A to B as tasks progress within each panel, and no switch from B to A. In addition, consistency requires players not to both create and destroy when they are richer (or poorer). The further right (left) off the origin a dot is, the more cumulated money a dictator is willing to give up in order to create (destroy) recipient's income when she is *richer* than the recipient. The further up (down) off the origin a dot is, the more cumulated money a dictator gives up to create (destroy) recipient's income when she is *poorer* than the recipient.

For example, imagine a GARP-consistent dictator, who switches from A to B in task 6 of panel 1 and, consequently, also chooses B in tasks 7 through 9. Thus, four B choices at a cost of 10 tokens each, so *40 tokens to create* income for the recipient when *being richer*, revealing the parameter $.19 \leq \rho \leq .24$. Moreover, this dictator switches from A to B in task 8 of panel 4 and, consequently, also chooses B in task 9. Thus, two B choices at a cost of 10 tokens each, so *20 tokens to destroy* income of the recipient when *being poorer*, revealing the parameter $-.23 \leq \sigma \leq -.17$. In the Cartesian type space, this dictator would be depicted in a point X that is 40 token units to the right of S, and 20 token units below S. Accordingly, the Cartesian distance from X to S is $40 + 20 = 60$.

Table A3: Logic of the decision tasks in the MDG

	(1) Ahead – Create $\pi_i^A=170$ $\pi_j^A=70$					**(2) Behind – Create** $\pi_i^A=110$ $\pi_j^A=120$				
Task	π_i^B	π_j^B	Δ_i	Δ_j	$\rho\geq$	π_i^B	π_j^B	Δ_i	Δ_j	$\sigma\geq$
1	160	82	*-10*	*12*	*0.45*	100	132	*-10*	*12*	*0.45*
2	160	84	*-10*	*14*	*0.42*	100	134	*-10*	*14*	*0.42*
3	160	88	*-10*	*18*	*0.36*	100	138	*-10*	*18*	*0.36*
4	160	94	*-10*	*24*	*0.29*	100	144	*-10*	*24*	*0.29*
5	160	102	*-10*	*32*	*0.24*	100	152	*-10*	*32*	*0.24*
6	160	112	*-10*	*42*	*0.19*	100	162	*-10*	*42*	*0.19*
7	160	124	*-10*	*54*	*0.16*	100	174	*-10*	*54*	*0.16*
8	160	138	*-10*	*68*	*0.13*	100	188	*-10*	*68*	*0.13*
9	160	154	*-10*	*84*	*0.11*	100	204	*-10*	*84*	*0.11*
	(3) Ahead – Destroy $\pi_i^A=140$ $\pi_j^A=130$					**(4) Behind – Destroy** $\pi_i^A=90$ $\pi_j^A=180$				
Task	π_i^B	π_j^B	Δ_i	Δ_j	$\rho\leq$	π_i^B	π_j^B	Δ_i	Δ_j	$\sigma\leq$
1	130	118	*-10*	*-12*	*-5.00*	80	168	*-10*	*-12*	*-5.00*
2	130	116	*-10*	*-14*	*-2.50*	80	166	*-10*	*-14*	*-2.50*
3	130	112	*-10*	*-18*	*-1.25*	80	162	*-10*	*-18*	*-1.25*
4	130	106	*-10*	*-24*	*-0.71*	80	156	*-10*	*-24*	*-0.71*
5	130	98	*-10*	*-32*	*-0.45*	80	148	*-10*	*-32*	*-0.45*
6	130	88	*-10*	*-42*	*-0.31*	80	138	*-10*	*-42*	*-0.31*
7	130	76	*-10*	*-54*	*-0.23*	80	126	*-10*	*-54*	*-0.23*
8	130	62	*-10*	*-68*	*-0.17*	80	112	*-10*	*-68*	*-0.17*
9	130	46	*-10*	*-84*	*-0.14*	80	96	*-10*	*-84*	*-0.14*

Note: To ensure that stakes are comparable across panels, every panel k has approximately (i.e. constrained on only using integers) the same mean pie size $\bar{P}_k = \frac{1}{18}\sum_{t=1}^{9}(\pi_{i,t}^A + \pi_{j,t}^A + \pi_{i,t}^B + \pi_{j,t}^B)$. Ahead-Create has 254 tokens, Ahead-Destroy 246, Behind-Create 244, and Behind-Destroy 246.

A.7 Elicitation of Moral Opinions

Aiming at eliciting people's home-grown *moral opinions* and not their perception of a *social norm*, I deliberately choose not to use the incentivization of Krupka and Weber (2013), which rewards people for correctly guessing what the majority believes to be appropriate behavior. However, since I do not provide monetary incentives for people to state the truth, it may be questionable whether people's answers $\hat{a}_i$ to the moral opinion question are actually a good measure of people's true moral opinions $\widetilde{a}_i$.

There are two distinct potential concerns: non-incentivized answers might be (i) noisy and (ii) biased. Whereas noisy answers would only reduce statistical power and make it more difficult to establish significant results, biased answers could possibly jeopardize the interpretation of the experimental results. Answers would be noisy if people cared *too little* about giving a truthful answer. Instead, answers would be biased if people cared *too much* about giving a specific untruthful answer. In particular, two potential biases come to mind:

On the one hand, knowing there is a 50% chance of being the dictator in the subsequent MDG, people may want to avoid self-commitment. Consequently, the stated $\hat{a}_i$ would be more

selfish than the true $\widetilde{a}_i$. As long as $\hat{a}_i = \tau a^* + (1-\tau)\widetilde{a}_i$ with $0 \leq \tau \leq 1$, i.e. as long as people declare it morally right to forgo material payoffs for the true cause but underreport the price they find morally right to pay for it, this bias would not constitute a problem for interpreting the data but just underestimate the distance between the true $\widetilde{a}_i$ and the observed choice a^{**}.

On the other hand, even in the absence of monetary incentivization to coordinate on a social norm, people may still *intrinsically* want to conform with a social norm. In that case, their answers to the moral opinion question could, potentially, not reveal their true $\widetilde{a}_i$ but rather their estimate of the socially expected answer. If subjects wanted their moral opinion to conform with the prevalent social norm, in *Introspect-i* they would have a handy device to align their behavior with the descriptive norm. The fact that moral opinions in *Introspect-u* and *Introspect-i* are virtually identical suggests that people's non-incentivized moral opinion reveals indeed (something close to) their true $\widetilde{a}_i$.