Egocentric Norm Adoption*

Thomas Neuber†

September 22, 2020

Job market paper
[click here for the latest version]
Preliminary draft; please do not circulate.

Abstract

Social norms pervade human behavior, but the importance of any norm is usually open to interpretation. This paper provides experimental evidence showing that people perceive the relevance of norms egocentrically: they believe in norms that are good for them in the sense that they would personally profit from these norms if everybody adhered to them. Many previous experiments have studied conflicting norms and have shown that people favor principles that allow them to be selfish. This paper also presents a (virtual) laboratory experiment in which subjects face conflicting norms. However, they decide over others instead of themselves, while their own payoffs depend on further participants' choices in the same decision contexts. Despite the absence of any personal motives, subjects deciding over others rely more intensively on normative principles that align with own their interests. Afterward, we ask subjects to predict the decisions of others, who might have different interests. Beliefs display the same kind of bias as found for decisions, pointing towards an unconscious mechanism. Survey answers provide further support for the importance of egocentrism: subjects who report pronounced perspective-taking are less biased.

Keywords: egocentrism, experiment, fairness, social norms
JEL Codes: C91, D63, D91, Z13

*I am thankful to Thomas Dohmen and Armin Falk for many discussions, to Jana Hofmeier, Axel Wogrolly, and Florian Zimmermann for helpful comments, and to Holger Gerhardt for support with the technical setup. The study was registered in the AEA RCT Registry under the unique identifying number *AEARCTR-0005774*. Funding by the German Research Foundation (DFG) through CRC TR 224 (Project A01) is gratefully acknowledged.

†Bonn Graduate School of Economics (BGSE), University of Bonn; thomas.neuber@uni-bonn.de

1 Introduction

When do people consider certain social norms to be important and adopt them as part of their behavior? Answering this question is instrumental in predicting choices, understanding how social norms evolve, and designing institutions. In fact, important social phenomena can often not be explained by pure self-interest, such as donations, volunteering, and honesty. Ample experimental evidence has confirmed the presence of these behaviors in non-strategic settings, such as giving in the dictator game, contributions in the public goods game, and honesty at a monetary cost (Fischbacher and Föllmi-Heusi, 2013). Social norms, defined by Ostrom (2000) as "shared understandings about actions that are obligatory, permitted, or forbidden" (pp. 143), are a widely used explanation for such behaviors (López-Pérez, 2008; Kessler and Leider, 2012; Kimbrough and Vostroknutov, 2016). The literature on *strong reciprocity* (Fehr and Gächter, 2000; Falk and Fischbacher, 2006) has extensively studied how people enforce social norms by rewarding behavior that complies with social norms and punishing behavior that violates them, even if this is costly (Fehr and Gächter, 2000). Other work has explored the precise nature of what people consider to be fair (Konow, 2003), e.g., contrasting egalitarians and libertarians (Cappelen et al., 2007). Our paper is concerned with *why* people follow certain norms. The leading explanation is evolution through selection by fitness, i.e., reproductive success (Becker, 1976; Güth and Kliemt, 1998), or imitation of successful others (Sugden, 1989; Richerson and Boyd, 2005). An example of such cultural evolution is that people today have less egalitarian gender norms if their ancestors traditionally used the plow, which facilitated labor division between women and men in agricultural societies (Alesina, Giuliano, and Nunn, 2013). Within the evolutionary process, individuals do not reason about the sense of norms. This paper proposes a complementary view by which people actively consider the implications of social norms. According to *egocentric norm adoption*, people judge the importance of a norm by its consequences and focus on their private outcomes. Of course, the consequences of a norm follow not only from a person's *own* behavior but, importantly, also from *others'* behavior. They are inclined to adopt a norm if the outcome that it implies is favorable and subconsciously base their judgment chiefly on the outcome for themselves. Thus, if a person wants others to follow a specific norm, simply because the consequences would be good for her, she will feel that the corresponding norm is objectively important and adhere to it herself. This paper presents an experiment with an identification strategy that exploits the role played by others and the presence of egocentrism. The design has the following features: First, subjects are affected by others' choices. Second, they also decide in the same decision contexts themselves but are not affected by their own decisions. Third, subjects' interests are exogenously varied, i.e., they are randomly allocated to roles that profit or lose from certain norms. Only if people judge the importance of norms by their consequences and

do so in an egocentric manner, subjects should make decisions over others that favor their own roles. We find clear evidence for such behavior.

In the experiment, subjects are randomly assigned to groups of two, and other subjects choose allocations of points for each group's members. The possible allocations involve tradeoffs between two different fairness norms, where each of the principles favors one of the group members. Subjects decide over the allocations in other groups along a circle: Group 1 decides over Group 2, Group 2 over Group 3, ..., and Group N over Group 1. Therefore, no subject can influence their own payoff. Egocentric norm adoption would predict that subjects favor the fairness principle from which they would profit themselves, despite holding no stakes in their own decisions. The experiment does not just consist of a single decision context but two. The *EF Procedure* trades-off equality against efficiency,[1] while the *EQ Procedure* involves equality and equity, i.e., the principle of accountability.[2] Subjects have distinct roles for each procedure that determine from which respective normative principle they profit, and the roles of subjects in adjoining groups are crossed. Consequently, each subject shares exactly one role with each player over whom she decides. This feature allows us to distinguish the context-specific effect proposed in this paper, whereby subjects' own interests matter, from any person-specific effects, like favoritism towards a specific player. The experiment's main result is that subjects' decisions over others are biased in favor of their own roles, thereby favoring one of the players in the EF Procedure and the other player in the EQ Procedure. After the subjects have decided, we elicit their beliefs about the choices of others, not conditioning on roles. Beliefs show very similar biases to those observed for decisions, suggesting that the main effect arises mostly unconsciously. As part of the questionnaires at the end of the experiment, we measure different aspects of empathy. In line with the interpretation of egocentrism driving the results, we find that decisions are less biased among subjects who report pronounced perspective-taking.

This paper's experiment provides evidence for egocentric norm adoption in the tradeoff between pairs of norms that are at conflict. A related paper by Hofmeier and Neuber (2019) is concerned with how the norm of helping is traded-off against material self-interest, depending on how much people require help themselves. In the experiment, *senders* can pay money to avoid that *receivers* have to eat different food items containing dried insects. They know what receivers would be willing to pay for themselves, which mutes the role of beliefs. All subjects act as senders but might be selected to act as receivers at the end of the experiment. The main result is that people pay more for others if they also pay more for themselves. This relationship holds between different subjects and also exists within individual subjects' decisions across different items (further insights

[1] Throughout the paper, we will denote the tradeoff between equality and efficiency as a *fairness* tradeoff, although efficiency in itself might not be considered a fairness criterion. However, efficiency is nonetheless relevant for fairness judgments (see Konow, 2001).

[2] For the empirical relevance of different fairness views, see Konow (2003) and Cappelen et al. (2007).

are discussed in Section 6). Subjects are thus imperfectly empathic, i.e., they act not only upon receivers' preferences but also upon their own. Egocentric norm adoption can explain this finding: people feel obliged to help others if they want help themselves. The experiments presented in the current paper and in Hofmeier and Neuber (2019) both stress the negative side of egocentric norm adoption, i.e., its egocentric aspect. Applied, e.g., to politics, the mechanism is a likely explanation for inefficient disagreement about fairness standards and distributive policies. However, egocentric norm adoption also has (perhaps primarily) a positive side, which is the aspect of norm adoption: people *do* think about how they would feel about their behavior themselves and therefore adhere to norms. In the experiment by Hofmeier and Neuber, this is indeed quite apparent: Many people are willing to give substantive amounts, just not optimally targeted at the receiver–item-combinations where the benefit for others would be largest. Egocentric norm adoption can thus have positive consequences in many social situations and, in particular, promote cooperation between individuals with shared interests. It can, e.g., motivate people to vote in large elections because they would like others who share their political preferences to do the same. More generally, egocentric norm adoption can help overcome collective action problems and supplying public goods because people in these situations depend on mutual help.[3]

The paper is related to multiple strands of literature that previously have been mostly unconnected. First, it is related to the literature on motivated reasoning and beliefs (Kunda, 1987, 1990; Oster, Shoulson, and Dorsey, 2013; Bénabou and Tirole, 2016). In particular, an extensive literature has been concerned with motivated beliefs in the domain of fairness. In an early contribution, Messick and Sentis (1979) find evidence for self-serving fairness views in a hypothetical setting regarding the remuneration for work conducted with another person who has worked for a longer or shorter time, respectively. In the economic literature, Konow (2000) elicits fairness views as real decisions over allocations between others. Konow shows that subjects who behaved unfairly due to selfish incentives subsequently adjust their fairness views and interprets this as proof for cognitive dissonance reduction (Festinger, 1957; Akerlof and Dickens, 1982).[4] Dana, Weber, and Kuang (2007) add "moral wiggle room" to the dictator game by reducing transparency and find decreased giving. Several further contributions have studied how people who are facing monetary incentives to behave unfairly exhibit more selfishness under circumstances which permit sustaining a positive self-image (Gino, Norton, and Weber, 2016). Among the identified kinds of "excuses" are competing (fairness) norms (Rodriguez-Lara and Moreno-Garrido, 2012; Bicchieri and Chavez, 2013; Barron, Stüber, and Veldhuizen, 2019; Kassas and Palma, 2019), sharing the benefits of unethical behavior with others

[3]This explanation is complementary to other contributing factors such as altruism (Becker, 1974), warm glow (Andreoni, 1990), and reciprocity (Fehr and Gächter, 2000; Falk and Fischbacher, 2006).

[4]However, Cerrone and Engel (2019) show that revealing one's fairness view is not sufficient to eliminate subsequent selfish behavior.

(Gino, Ayal, and Ariely, 2013), possible misdemeanor of those to be treated unfairly (Di Tella et al., 2015), ambiguity or risk over the efficacy of prosocial behavior (Haisley and Weber, 2010; Exley, 2016), and supposed mistakes in decision making (Exley and Kessler, 2019). In all of these contributions, the provision of direct monetary incentives biases fairness views. Self-serving fairness views have also been documented in bargaining contexts, contributing to bargaining impasse between parties who do not sufficiently appreciate the other side's arguments (Thompson and Loewenstein, 1992; Loewenstein et al., 1993; Babcock et al., 1995; Babcock and Loewenstein, 1997; for a successful replication, see Hippel and Hoeppner, 2019). This bias is in line with research showing that people who successfully convince themselves of a particular argument in their favor are better at convincing others (Smith, Trivers, and Hippel, 2017; Schwardmann and Weele, 2019), for which Schwardmann, Tripodi, and Weele (2019) provide additional evidence in the field setting of a debating competition.[5] Our paper contributes to the above literature by demonstrating bias in a context without any motives that would conflict with objective fairness. In the experiment, subjects do not need to legitimize any past actions, their decisions do not affect their payoffs, and they do not need to be convincing. Instead, a given subject could do what she objectively believes to be fair and—maybe—hope that others disagree with her view, thereby allocating more points to her than her own decisions would imply. The subject could even think that receiving more points than she would allocate to someone in her position would happen to be a fair outcome, perhaps because she feels especially deserving as a person or is in particular need of money. The observed bias is evidence that such reasoning is not the whole story. Epley and Caruso (2004) have suggested that people are convinced of self-serving ethical judgments as a result of egocentrically biased affective reactions (see Zajonc, 1980; Haidt, 2001; Slovic et al., 2002) that are automatic and unconscious.[6] This paper agrees and shows that egocentric perceptions of potential outcomes do not just affect how people feel about narrow situations that involve themselves. Instead, egocentrism also translates into people's actions and how they treat others, apparently because it alters different norms' perceived importance. The experiment thereby shows that egocentrism can have consequences in situations where people could genuinely claim that they are free from any "conflict of

[5]Concerning the mechanism behind self-persuasion, Babcock et al. (1995) show that the egocentric bias in fairness views is reduced to statistical insignificance when subjects only learn about their roles after having read the instructions, i.e., self-persuasion seems to work through differential information encoding. Similarly, in the context of self-interested financial advice, Gneezy et al. (2020) show that self-deception about the truly best options is more pronounced when advisors know about the selfish incentives already before they make their private evaluations. Zimmermann (2020) empirically shows that another mechanism to arrive at motivated beliefs is selective memory. The findings show that creating and sustaining motivated beliefs is an active mental process.

[6]Regarding the aspect of unconsciousness, a psychological literature has been concerned with how judgments regarding, e.g., the quality of an applicant, can be "contaminated" by affective reactions (Wilson and Brekke, 1994), finding that people's awareness of their internal processes is insufficient to overcome the resulting biases. Relatedly, Bocian and Wojciszke (2014) show that others' immoral behavior is judged less harshly by observers if the latter themselves profited from the behavior.

interest." A judge in a court case, e.g., might not know either of the involved parties, not hold any private interests in the matter under review, and not be prejudiced. Maybe, however, the judge shares certain case-relevant features with one of the parties (e.g., being female or male in the context of gender discrimination). In light of this paper's findings, the judge's decision could still be biased.

The paper is thus also related to a second strand of literature concerned with in-group–out-group bias. This research area started from the observation that experimental subjects tend to favor other subjects from their own group over subjects from other groups even when the criteria used to form groups are "minimal" (Tajfel, Billig, and Bundy, 1971; Billig and Tajfel, 1973). This finding is now commonly explained with social identity theory (SIT; Turner, Brown, and Tajfel, 1979). The latter starts from the premise that part of individuals' identity is their social identity, which they derive from group memberships. People increase their self-esteem by adopting more favorable beliefs about in-group members than out-group members, as evident in ratings (Mullen, Brown, and Smith, 1992), and treating the former better than the latter. Owing to the observations that individuals usually belong to many social groups and that those groups typically overlap, there is an interest in the effect of crossing group categorizations between individuals (Brown and Turner, 1979), i.e., the relations between in-groups, single out-groups, and double-outgroups. An additive pattern seems to prevail: in evaluations, people behave as if they count the number of dimensions in which another person belongs to their in-group and subtract the number of out-groups to which the same person belongs (Crisp and Hewstone, 1999). Chen and Li (2009) examine the effects of minimal groups within the setting of commonly used paradigms of experimental economics. They find that, relative to out-group members, members of a subject's in-group experience more altruism, increased positive reciprocity, and decreased negative reciprocity. Our paper relates to this literature in that egocentric norm adoption naturally gives rise to a phenomenon akin to in-group–out-group bias: people favor others with whom they share the same economic interests, i.e., discrimination arises between *interest* groups. The experiment rules out classical in-group–out-group bias by ensuring that subjects in adjoining groups always share precisely one role. Therefore, both group members for whom a player chooses an allocation are in one of her in-groups and one of her out-groups, such that SIT would not make any prediction for differential treatment. The crossing of roles also implies that egocentric norm adoption favors a different participant for each of the two decisions that a subject takes.

Finally, the present research is related to a mostly theoretical literature on "Kantian" behavior, which proposes that human behavior is following a version of Kant's categorical imperative to "[a]ct only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant, 1996, p. 73). Loosely speaking, the economic literature says that a subject has Kantian moral concerns if she prefers using

strategies that would benefit her also if everyone else also adopted. Roemer (2010) shows that in the presence of externalities, equilibria arising from Kantian maximization dominate Nash equilibria. Alger and Weibull (2013) show that under assortative matching of individuals who interact, evolution should converge to a mixture of selfish and Kantian preferences.[7] Leeuwen, Alger, and Weibull (2019) empirically investigate the presence of deontological preferences. They do so by letting subjects play both roles in different two-player dilemmas, eliciting their beliefs about others' strategies, and structurally estimating subjects' preferences. Intuitively, Kantian preferences predict strategies that would work especially well if subjects played with themselves in different roles. In the sequential prisoner's dilemma, e.g., those cooperating as the first mover also tend to cooperate with a higher probability as the second mover.[8] As has been shown by Blanco et al. (2014), this correlation can, to a large extent, be explained by beliefs about others' behavior, i.e., by false consensus, but not entirely. Since there is no experimental treatment involved, several different preference-based explanations for this finding are possible (see Blanco et al., 2014). A latent class analysis conducted by Leeuwen, Alger, and Weibull (2019) indicates that deontological preferences do well in explaining the observed patterns. Like the literature on Kantian behavior, this paper proposes that people mainly care about their own outcomes and exhibit rule-based behavior. Conceptually, we bridge the above literature to the much larger literature on social norms, an obvious ingredient of rule-based behavior. Moreover, we suggest that the process of selecting behavioral rules is not highly abstract, as the reference to Immanuel Kant would suggest, but mainly unconscious, which is confirmed by our finding of biased beliefs. Empirically, our identification strategy does not exploit patterns within the strategies that subjects choose but uses the aspect of egocentrism, which has played no role in previous experiments. Since roles are assigned exogenously, egocentric norm adoption is cleanly identified. The results from our experiment show that egocentrism plays a vital role in how people select behavioral rules. This property is clearly opposed to the idea of deontological ethics but apparently a more realistic characterization of people's intuitive behavior.

The remainder of this paper is organized as follows: Section 2 introduces the design in detail. The derivation of the hypotheses follows in Section 3. Section 4 presents the main results. Subsequently, Section 5 conducts an analysis of heterogeneity in the observed effects. Finally, Section 6 summarizes the paper and provides a discussion of other contexts in which egocentric norm adoption might be important.

[7] See also Bergstrom (1995) for an early contribution and Alger and Weibull (2019) for a review.

[8] A similar approach is used by Costa-Gomes, Ju, and Li (2019), who find what they call "role-reversal consistency."

2 Experiment

The experiment's design is such that subjects' self-interest is aligned or opposed to certain kinds of choices in decision contexts that subjects are facing themselves. However, their *own decisions* are entirely irrelevant for their payoffs—a feature that was made highly salient throughout the experiment. The experiment's structure also ensures that subjects' choices do not affect those players on whom their own payoffs depend, avoiding considerations of reciprocity (Fehr and Gächter, 2000; Falk and Fischbacher, 2006). Interests in decision contexts necessarily give rise to groups or, more specifically, *interest groups*. The experimental design comprises two different procedures, such that each player has two roles. Own roles and roles of subjects for whom players decided are crossed, such that subjects share exactly one role with each other subject over whom they decide. This allows differentiating the effect of egocentric norm adoption from group favoritism in the sense of social identity theory. Groups in terms of identity would take into account both roles of subjects. If both procedures were equally relevant for subjects' identities, social identification would predict no effect. If one procedure was more important than the other, social identification would predict that choices by a given player favor the same subject in both decisions. In contrast, egocentric norm adoption predicts that a different subject is favored in both procedures, i.e., always the one sharing the decider's role in the respective procedure. Subjects had to answer a number of control questions correctly to make sure that they had understood the structure of the experiment.

2.1 Design

Experimental sessions are run with a multiple of four participants. The participants, or players, are randomly allocated to groups of two, which are numbered consecutively from 1 to N. In each group, one participant is called *Player X* and the other *Player Y*. All participants receive a fixed participation fee of €4 and, during the experiment, points that are each worth €0.01. Importantly, no player makes any decision regarding their own group. Instead, groups decide for players in other groups along a circle, i.e., Group 1 decides for Group 2, Group 2 decides for Group 3, ..., and Group N decides for Group 1. Players are being told so before learning any of the details of the experiment. Every player makes two decisions, each between 20 different options. One decision is about the tradeoff between equality and efficiency; the other is about the tradeoff between equality and equity, i.e., attribution of responsibility. For the EF Procedure, players in each group are split between two roles, of which one profits from efficiency and the other from equality. For the exposition in this paper, we denote the former by A and the latter by B. For the EQ Procedure, roles are denoted by a and b, where Role a profits from equity and Role b from equality. The names of roles are not used in the instructions, and the same is true for the two procedures. Their order is randomized on the player-level, and they are simply

referred to by their order of appearance as "Procedure 1" and "Procedure 2". Any two players in any two adjoining groups share exactly one role. Denoting players by tuples of roles, the structure can thus be represented as in Figure 1. This structure ensures

$$\cdots \Rightarrow \begin{matrix}(A,a)\\(B,b)\end{matrix} \Rightarrow \begin{matrix}(A,b)\\(B,a)\end{matrix} \Rightarrow \begin{matrix}(A,a)\\(B,b)\end{matrix} \Rightarrow \begin{matrix}(A,b)\\(B,a)\end{matrix} \Rightarrow \cdots$$

Figure 1: Structure of the Experiment

that no player is ever confronted with another player belonging to a "double out-group" in terms of their roles and thereby alleviates concerns about in-group–out-group bias.[9] Players first learn about types in their own groups and potential payoff consequences for themselves and their partner in the group before being informed about the details of the succeeding group, for which they decide. At the end of the experiment, one of the two procedures is selected at random. The respective decision of one player from either all odd- or all even-numbered groups is implemented. These players themselves receive 1,000 points, which is the maximum that can be attained in the experiment.

Before subjects are made familiar with any further details of the experiment, they are asked to complete the following task that is explained to them.

Estimation Task On their computer screens, subjects see a three-second countdown, after which they are shown an image for the duration of two seconds. It shows a number of blue dots on a yellow background. Immediately after the image has again disappeared, subjects are forwarded to another page on which they have 15 seconds to enter an estimate for the number of dots that they saw. Their task is to minimize the absolute difference between their estimate and the true number of dots.[10] The instructions state that a better estimate will increase their chances of receiving additional money during the experiment. Before completing the actual task, subjects engage in a trial task that is identical to the incentivized task except for the number of dots that are shown. The respective images that subjects see are identical for all participants, showing 40 dots for the trial task and 53 for the one that is potentially payoff-relevant.

[9] There exists a literature on the effect of multiple and crossed categorizations on intergroup behavior (see, e.g., Brown and Turner, 1979; Vanbeselaere, 2000). This literature is, however, not concerned with decisions that are specific to any particular characteristics (in our case roles) but rather with how differences are aggregated. In our experiment, subjects always share exactly one of two roles with participants for whom they make decisions.

[10] The task of estimating the number of dots is inspired by the one used in Fliessbach et al. (2007). The original task, however, asks subjects to make the binary judgment of whether the number of dots was higher or lower than a given integer. Asking for a specific estimate instead allows for a more fine-grained assessment of performance and thereby helps to avoid ties between pairs of subjects. Tajfel, Billig, and Bundy (1971) use a similar setup to induce groups in terms of subjects under- or overestimating numbers of dots. In contrast, subjects in this experiment do not learn anything about their own performance (until the very end of the session), and thus no groups are induced by the estimation task.

After having completed the estimation task, subjects are introduced to the basic setup of the experiment, i.e., the circular structure. They are informed about the group that they have been allotted to, their name (X or Y), and the name of their partner. It is clearly spelled out who makes decisions regarding the group to which they belong themselves and for which group they are asked to make decisions. They are explicitly told within a highlighted box that they will in no way be able to influence the allocation of points within their own group. The two procedures are introduced as "Procedure 1" and "Procedure 2", and the details on them follow next in the respective order of their names.

Efficiency (EF) Procedure The Efficiency (EF) procedure concerns the tradeoff between equality in points that are allocated to both players of a group vs. efficiency in terms of total points for both group members. The possible allocations of points are shown in Table 1.

Table 1: Payoffs for the Efficiency Procedure

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	200	300	385	460	525	585	640	690	735	775	811	843	871	896	918	937	953	967	979	990
B	200	190	180	170	160	150	140	130	120	110	100	90	80	70	60	50	40	30	20	10
$\sum$	400	490	565	630	685	735	780	820	855	885	911	933	951	966	978	987	993	997	999	1,000

Columns show the 20 options among which subjects can choose for their respective succeeding groups. The row below the option numbers shows the points that the player in Role A receives as part of each allocation. This number is strictly increasing over options but in decreasing increments, i.e., the number of points mimics a strictly concave function. Increases start at 100 points and decrease to a minimum of eleven points. The number of points that the player in Role B receives equals that of the other player only for the first option and then decreases in constant increments from 200 down to 10. The bottom row shows the total number of points, which ranges from 400 to 1,000. Relative to the fully equal outcome therefore, efficiency can be increased by a factor of up 2.5. However, efficiency gains decrease from 90 points going from Option 1 to Option 2 to just a one-point difference between options 19 and 20. Thus, going from lower to higher options, inequality increases at diminishing returns in terms of efficiency.

Equity (EQ) Procedure At the very beginning of the experiment, all players have engaged in an estimation task, which they were told increased their chances of getting additional money (see above). The facts that subjects cannot learn about their performance and that everybody took part in the same task under the same conditions mutes any motives regarding self-esteem. The estimates that subjects gave are used for the Equity (EQ) procedure in which, however, only the estimate of the player in Role a is compared to the estimate of another player from a non-adjoining group. If the estimate

of the player in Role a was better than the other estimate, the group receives 1,000 points and otherwise it receives no points. The comparison does not affect the payoffs of players in any other groups (the other player whose estimate enters the comparison has Role b in their own group). Conditional on the player in Role a having secured the points, one allocation needs to be chosen from the 20 options provided in Table 2.

Table 2: Payoffs for the Equity Procedure

#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
a	500	525	550	575	600	625	650	675	700	725	750	775	800	825	850	875	900	925	950	975
b	500	475	450	425	400	375	350	325	300	275	250	225	200	175	150	125	100	75	50	25

As for the EF procedure, Option 1 implements equality of points between roles, i.e., players. For every further option, 25 points are added for the player in Role a (who secured the points) and the same number of points is deducted from the player in Role b (whose performance is irrelevant for the group). Thus higher-numbered options constitute allocations which reflect accountability for the total points that the group received, i.e., a reward for the player who won the points.

Subjects are made familiar with the procedures first in the context of their own group. This is similar to many real-life situations in which people know about their interests (e.g., that they are rich or poor) before considering a particular decision problem (voting in favor of a redistributive policy proposal or against it). The depiction of potential payoffs mimics Table 1, except that participants do not see the names of the roles but those of the respective participants in their group. The row for Player X is always shown on top and that for Player Y below, i.e., the two rows might be reversed.[11] Everything is explained in detail, and it is made salient throughout that the decision rests with the members of the sending group. After reading through the instructions for the two procedures, subjects arrive at a screen with a number of control questions. These first make sure that subjects have understood the structure of the experiment, i.e., who decides for whom and whose estimate counts for Procedure EQ. Other questions make sure that subjects interpret the choice tables correctly. Subjects were able to go back to the previous screen with the instructions and reread them to facilitate answering the questions. In case of problems, they could approach the experimenter, who was available via phone, text message, or email.[12]

After the successful completion of the first set of control questions, participants read about the respective groups for whom they would be asked to make decisions. They are informed about the names of players (again X and Y) and how the roles in this other

[11] Since players' names are independent or their roles, no bias can be introduced here. However, the fixed and transparent order facilitates understanding.

[12] Using an online conference platform instead would not have allowed for one-to-one communication between subjects and the experimenter.

group compare to the roles of themselves and the other player in their own group. They then proceeded to two screens mimicking almost exactly the screens they had just seen before, again explaining the two procedures, but this time adapted to the group for which they decide. They then answer a second set of control questions that make sure subjects interpret the tables correctly.

Lastly, before the two decisions are made, it is explained to subjects in detail how the actual payoffs at the end of the experiment will be determined. In particular, with a chance of 50%, a decision by a subject from the proceeding group is going to be implemented for their own group. With 25% probability, a decision of them is implemented for the succeeding group, in which case they themselves receive 1,000 points.[13] And, with a chance of 25% percent, a decision by their group partner is implemented, and their own payoff is determined by another task that is independent of their own decisions (see the paragraph on belief elicitation below). This screen is followed by a third and final set of control questions that make sure that subjects remember everybody's roles, ascertaining that the crossing of roles is apparent to subjects. Next come the decisions for the respective succeeding group, one after the other in the subjective-specific order. It is one again spelled out for which group the decision is made, and no option is preselected.

Belief Elicitation After having made their two decisions, players' beliefs about choices by others are elicited. Specifically, they are asked to guess the average choice of that subjects from other groups have made for groups that, in terms of their role compositions, are identical to the one for which they have decided themselves. The accuracy of their guess determines their payoff in case a decision of the other player in their own group is implemented for the following group, i.e., with a probability of 25%. For odd-numbered (even-numbered) groups, the true average choices for the true procedures are calculated over all subjects within the same session who were in the other odd-numbered (even-numbered) groups. This makes sure that the selected other subjects decided for groups that, abstracting from players' names (who was X and who was Y), were identical to the one for which the respective participant was deciding herself. The subject then receives 500 points if her guess is exactly correct and 250 points as long as the correct answer falls into the range of the five options closest to their guess. The predictions are made on tables that look exactly like the ones for the decisions, and the range of options for which the selected option still implies 250 points, which towards the ends of the table becomes

[13]For subjects whose decisions are implemented, the compensation is thus fixed and thereby independent of their roles. Moreover, the number of points that deciding subjects receive (1,000) is always larger than the payoff for any of the two subjects over whom they decide. These properties of the experimental design alleviate concerns that subjects' decisions over others might be influenced by expectations about their own payoffs, e.g., due to inequality aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). Also note that if subjects in Roles B and b should choose more equal options because they wanted to reduce the gap to subjects in Roles A and a (in the succeeding group) *in expectations*, a negative correlation should be observed between choices and beliefs. As Section 4.2 will show, the opposite is the case.

asymmetric, is highlighted.

The experiment concludes with a survey in which subjects are asked about basic socio-demographic characteristics like age, gender, and their studies, and participants complete several questionnaires. The details with the corresponding results are presented in Section 5.

2.2 Procedure

The experiment was run starting from May 13 until May 20, 2020, and implemented as a *virtual lab experiment*. A total of 372 participants completed the experiment.[14] As in online experiments, subjects participated via the Internet. The experiment was programmed using *oTree* (Chen, Schonger, and Wickens, 2016), such that subjects could access it through their web browser using their own devices.[15] Since the experiment was run during the first phase of the COVID-19 pandemic, subjects presumably participated from home (the university library, e.g., was closed at the time). On the other hand, contrary to typical online experiment and just as in a usual laboratory experiment, subjects attended specific experimental sessions, i.e., they had to take part in the experiment at a pre-specified time and date, other participants in the respective same session were taking part simultaneously, and an experimenter was available to answer questions. On the first screen of the experiment, subjects were given contact details which they could use in case of questions. The experimenter was available via email, telephone, or text. The contact details had also been included in the automatized email communications with the subjects prior to the experiment (the invitation, an email with the personal link, and a reminder). Subjects made use of all contact methods. Participants were recruited from the subject pool of the *BonnEconLab* using the software *hroot* (Bock, Baetge, and Nicklisch, 2014). They received individual links, such that multiple effectively impossible for any subject to participate more than once. Participants were mostly university students, and the majority of subjects were women. For details of the sample composition, see Table B.1 (see the appendix).

3 Hypotheses

The main hypothesis of the paper is that participants make decisions favoring the role that they have been assigned to themselves for the respective procedure. To understand

[14] Four participants were excluded because they either stopped working on the experiment or because they could not answer the control questions.

[15] The invitation stated that subjects were required to use a regular desktop or laptop computer. In principle, however, the experiment was also fully functional on smaller devices such as smartphones or tablets.

the reasoning behind this conjectured effect in the absence of material incentives or, in fact, *any* motives, we develop a simple formal framework that attributes biased fairness views not to motivated cognition but to the (partial) inability to abstract from one's own role. From the gained insights, the paper's hypotheses are derived.

3.1 Formal Framework

The starting point is that, while considering the possible choice options, an agent experiences an affective reaction that is determined by her fairness views but also by the payoff implied for her own relevant role, because her perspective is inherently subjective. Her fairness views and level of subjectivity are, however, unknown to the agent. Instead, she simply knows what "feels best" for her. When being confronted with the choice that she has to make over others, she takes this task seriously and tries to empathize with those who will be affected.[16] She thus engages in the underlying normative trade-off and tries to learn about the importance of the involved norms. For this, she uses her affective reactions and asks how they came about. If she is perfectly capable of perspective-taking, she fully realizes the extent of subjectivity underlying her reactions, backs-out her true fairness-views, and takes an unbiased decision. If, however, she is affected by some degree of egocentrism, i.e., her ability of perspective-taking is imperfect, she underestimates the influence of subjectivity and arrives at fairness-views that depend on her own roles and at corresponding choices that are egocentrically biased.

3.1.1 Basic Setup

Formally, when considering a given option for one of the procedures, an agent experiences an affective reaction which is determined by her fairness views but in part also by the implied payoff for her own role.

$$\begin{aligned} React_{EF}(c_{EF}) &= \alpha\, Pay(c_{EF}, role_{EF}) - \beta_1\, Ineff(c_{EF}) - Inequal_{EF}(c_{EF}) \\ React_{EQ}(c_{EQ}) &= \alpha\, Pay(c_{EQ}, role_{EQ}) - \beta_2\, Unfair(c_{EQ}) - Inequal_{EQ}(c_{EQ}) \end{aligned} \tag{1}$$

To make the exposition easier, we assume that sets of possible values of options c_{EF} and c_{EQ} are the interval $[1, 20]$, i.e., the agent can choose intermediate options. The affective reaction to option c_{EF} for the EF Procedure, $React_{EF}$, comprises the respective own payoff Pay for the subjects relevant role $role_{EF} \in \{A, B\}$, the extent to which the respective option violates the equality norm, $Inequal_{EF}$, and the degree of implied inefficiency, $Ineff$. The reaction to option c_{EQ} for the EQ Procedure is also determined by the payoff for the subject's role and the violation of the equality norm and, in this case, by the violation of the fairness norm by which only the player should profit who secured the points. The

[16] If she did not care about the people over whom she decides, she would simply choose randomly.

influence of the payoff for the own role is determined by the level of subjectivity $\alpha \geq 0$ and the relative weights attached to the efficiency and the fairness norms are $\beta_1 > 0$ and $\beta_2 > 0$, respectively. For Roles A and a, the function *Pay* is strictly increasing while, for Options B and b, it is strictly decreasing. Thus, it may simply correspond to numbers of points. *Ineff* and *Unfair* are both strictly decreasing and strictly convex, as higher options are (decreasingly) more efficient or allocate more points to the responsible player, respectively. On the other hand, $Inequal_{EF}$ and $Inequal_{EQ}$ are both strictly increasing and convex, as higher options imply increasingly unequal payoffs for players. Moreover, we assume that all of the functions are differentiable.

The agent's intuitive reactions are thus best for some options $\tilde{c}_{EF}, \tilde{c}_{EQ} \in (1, 20)$. The agent knows how her reactions came about up to the three parameters α, β, γ_1, and γ_2. In scrutinizing the reasons for her reactions, she forms beliefs $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\gamma}_1$, and $\tilde{\gamma}_2$ about the parameters which, assuming an interior solution, must obey the two first order conditions.

$$\tilde{\alpha}\, Pay'(\tilde{c}_{EF}, role_{EF}) - \tilde{\beta}_1\, Ineff'(\tilde{c}_{EF}) - Inequal'_{EF}(\tilde{c}_{EF}) = 0$$
$$\tilde{\alpha}\, Pay'(\tilde{c}_{EQ}, role_{EQ}) - \tilde{\beta}_2\, Unfair'(\tilde{c}_{EQ}) - Inequal'_{EQ}(\tilde{c}_{EQ}) = 0$$

Clearly, the set of solutions is not atomic, and the intuitive optimum can be rationalized by different combinations of parameters. For example, a high value of r^*_{EF} for an agent in Role A could be due to strong subjectivity (large α) or due to strong efficiency concerns (large β_1). The agent starts her inference from prior beliefs about the true parameter values that follow independent Normal distributions with standard deviations of one. For $\tilde{\beta}_1$ and $\tilde{\beta}_2$, the means are the respective true values, while for $\tilde{\alpha}$, the mean is attenuated by $e \in [0, 1]$, i.e., the inability to abstract from one's personal vantage point.[17] Thus, the prior belief is given by $\mathcal{N}((1 - e)\,\alpha, 1)$. Her beliefs are the values that are most likely given her prior beliefs and the two first-oder conditions.

Lemma 1. *Assume that $\alpha, e > 0$. Then:*

1. *The agent underestimates her level of subjectivity, i.e., $\tilde{\alpha} < \alpha$.*
2. *The agent's beliefs about her true fairness preferences are egocentrically biased.*
   (a) *If she is in Role A, $\tilde{\beta}_1 > \beta_1$. Otherwise, i.e., if she is in Role B, $\tilde{\beta}_1 < \beta_1$.*
   (b) *If she is in Role a, $\tilde{\beta}_2 > \beta_2$. Otherwise, i.e., if she is in Role b, $\tilde{\beta}_2 < \beta_2$.*

Proof in Appendix A.1. □

[17] One could also interpret this assumption in the sense of cognitive dissonance. Subjects would then find it implausible that a wedge exists between their affective reactions and their true fairness judgments. However, the results on perspective-taking in Section 5 make ability/egocentrism seem more the more plausible interpretation.

3.1.2 Combining the Procedures

The above framework considers the decisions according to the two procedures independently, which suffices for the main predictions. Note, however, that both the Efficiency Procedure and the Equity Procedure involve equality as an overlap in the involved fairness norms, aligned with the interest of roles B and b, respectively. Thus, a participant with roles (B, b) always profits from equality, while one with roles (B, a) or (A, b) profits from equality according to one procedure and loses in the other. Lastly, the private interest of a participant with roles (A, a) is always opposed to equality. Using this feature, the setup might allow for insights into how egocentrically adopted norms can spill over from their source to other contexts. Formally, let us modify Equation 1 in the following way:

$$\begin{aligned} React_{EF}(c_{EF}) &= \alpha\, Pay(c_{EF}, role_{EF}) - \beta_1\, Ineff(c_{EF}) - \gamma\, Inequal_{EF}(c_{EF}) \\ React_{EQ}(c_{EQ}) &= \alpha\, Pay(c_{EQ}, role_{EQ}) - \beta_2\, Unfair(c_{EQ}) - \gamma\, Inequal_{EQ}(c_{EQ}) \end{aligned} \tag{2}$$

As before, the agent knows her true reaction functions up to the now four parameters α, β_1, β_2, and γ. All beliefs are the same as before, and the belief about γ is also follows a Normal distribution with a standard deviation of one, centered around the true value. From these modified assumption follow the below results.

Lemma 2. *Assume positive subjectivity and egocentrism ($\alpha, e > 0$). Then:*

1. *The agent underestimates her level of subjectivity, i.e., $\tilde{\alpha} < \alpha$.*
2. *The agent's beliefs about her true fairness-views are egocentrically biased.*
   - *(a) For roles A and a, it holds that $\tilde{\gamma} < \gamma$. Moreover, $\tilde{\beta}_1 > \beta_1$ and/or $\tilde{\beta}_2 > \beta_2$.*
   - *(b) For roles B and b, it holds that $\tilde{\gamma} > \gamma$. Moreover, $\tilde{\beta}_1 < \beta_1$ and/or $\tilde{\beta}_2 < \beta_2$.*
   - *(c) For roles A and b, it holds that $\tilde{\beta}_1 > \beta_1$ and $\tilde{\beta}_2 < \beta_2$.*
   - *(d) For roles B and a, it holds that $\tilde{\beta}_1 < \beta_1$ and $\tilde{\beta}_2 > \beta_2$.*

Proof in Appendix A.1. □

3.2 Predictions

In making her decisions, the agent tries to be impartial and therefore omits considerations regarding her own role. Using the basic setup of Section 3.1.1, the objective functions that she *wants* to maximize are thus the following.

$$\begin{aligned} u_{EF}(c_{EF}) &= -\beta_1\, Ineff(c_{EF}) - Inequal_{EF}(c_{EF}) \\ u_{EQ}(c_{EQ}) &= -\beta_2\, Unfair(c_{EQ}) - Inequal_{EQ}(c_{EQ}) \end{aligned} \tag{3}$$

In the objective functions that the she *actually* maximizes, however, the unknown parameters β_1 and β_2 are substituted by the agent's egocentrically biased beliefs $\tilde{\beta}_1$ and $\tilde{\beta}_2$, respectively. We again assume interior solutions and use that, by the assumptions from Section 3.1.1 the utility functions is concave. Under these conditions, the agent's choices c^*_{EF} and c^*_{EQ} are uniquely identified by the following first-order conditions.

$$\begin{aligned} -\tilde{\beta}_1 \, \mathit{Ineff}'(c^*_{EF}) - \mathit{Inequal}'_{EF}(c^*_{EF}) &= 0 \\ -\tilde{\beta}_2 \, \mathit{Unfair}'\big(c^*_{EQ}\big) - \mathit{Inequal}'_{EQ}\big(c^*_{EQ}\big) &= 0 \end{aligned}$$

Both optima are strictly increasing in the values of $\tilde{\beta}_1$ and $\tilde{\beta}_2$. In conjunction with the egocentric biases shown in Lemma 1, this leads to the main hypothesis of the paper.

Hypothesis 1. *For both procedures, subjects make choices favoring their own respective role.*

To test the hypothesis formally, denote by $i \in \{X, Y\}$ a subject in group $g \in \{1, 2, \ldots, N\}$. Within her group, the subject is randomly allocated to a role $r^{EF}_{i,g} \in \{A, B\}$ for the Efficiency (EF) Procedure and a role $r^{EQ}_{i,g} \in \{a, b\}$ for the Equity (EQ) Procedure. The subject's choice for Procedure EF is denoted by $c^{EF}_{i,g}$ and the one for Procedure EQ by $c^{EQ}_{i,g}$. Hypothesis 1 was preregistered, and in the pre-analysis plan we committed to running the two following regressions:

$$\begin{aligned} c^{EF}_{i,g} &= \delta_0 + \delta_1 1_A\big(r^{EF}_{i,g}\big) + \epsilon_{i,g} \\ c^{EQ}_{i,g} &= \zeta_0 + \zeta_1 1_a\big(r^{EQ}_{i,g}\big) + \eta_{i,g} \end{aligned} \tag{4}$$

The terms $1_A\big(r^{EF}_{i,g}\big)$ and $1_a\big(r^{EQ}_{i,g}\big)$ denote indicator functions for roles A and a, respectively. Since subjects in roles A and a would profit from higher-ordered choices made by the respective sending group, egocentric norm adoption predicts that both β_1 and γ_1 should be positive. Due to the crossing of types (tuples of players' roles) across adjoining groups, group identification could account for at most one positive coefficient. Therefore, the following hypothesis test will be conducted:

$$\begin{aligned} H_0: &\quad \delta_1 \leq 0 \vee \delta_1 \leq 0 \\ H_1: &\quad \zeta_1 > 0 \wedge \zeta_1 > 0 \end{aligned}$$

Rejection of H_0 would provide evidence for the presence of egocentric norm adoption. Note that H_0 includes a logical disjunction, which implies that the appropriate p-value for the test over both coefficients must be (weakly) larger than any of the two one-sided p-values referring to the individual coefficients. Appendix A.2 shows that an upper bound for the joint p-value is provided by the sum of the two separate one-sided p-values or, assuming symmetry, the average of the two-sided ones.

Equivalent testing procedures can also be applied to the two further hypotheses to be derived in this section. As has been shown in Lemma 2, an agent who loses from equality in both procedures (whose roles are A and a, respectively) and hence initially feels attracted to high choice options will view this as strong evidence for little concern with equality and for heightened concern also with at least one of the other norms. The converse is, of course, true for an agent with roles B and b. On the other hand, agents who profit from equality according to one procedure and lose from it in the other will notice that their initially preferred choices are somewhat contradictory as one seems to reflect strong concern about equality while the other does not. This leads to the last hypothesis.

Hypothesis 2. *Among participants with types whose private interests are aligned with or opposed to equality for* both *procedures, the effect of their own roles is larger than among other participants.*

In other words, we thus expect spillovers of roles to the respective other decision context, i.e., a positive effect of Role A on the decision according to the EQ Procedure and, similarly, a positive effect of Role a on the choice for Procedure EF.

In the formal framework that was introduced here, the bias in choices arises unconsciously and, as becomes clear by Lemma 1, is accompanied by distorted beliefs about fairness. Research on the *false-consensus effect* has shown that people typically overestimate the extent to which others share their views, which in the context of this experiment would mean that they project their own bias upon others. This leads to a further hypothesis.

Hypothesis 3. *Predictions about the choices made by others co-move with subjects' own choices and reflect egocentric norm adoption.*

Since people probably do not fully project their own views upon others but will moderate their predictions to some degree, the effects for beliefs should be expected to be a bit smaller than those for the respective decisions.

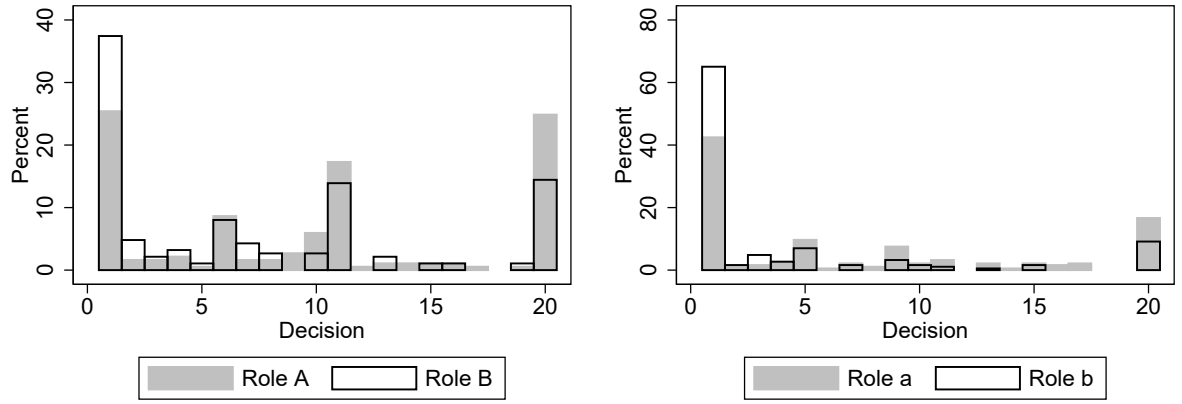
4 Main Results

This section presents the main results of the experiment, starting with the decisions in Section 4.1 and proceeding with the analysis of beliefs in Section 4.2.

4.1 Decisions

The decisions that subjects made in the experiment are visualized in Figure 2. The two panels are identically constructed, with Figure 2a on the left displaying the distribution of decisions for the EF Procedure and Figure 2b on the right showing the decisions for

the EQ Procedure. In displaying the distributions of decisions, the panels differentiate between the two roles that are relevant for the respective procedure: for the EF procedure, these are Role A (shaded), profiting from efficiency, and Role B (light), profiting from equality; and for the EQ procedure, the relevant roles are a, which is favored by the equity principle, and b, again benefiting from equality. For both procedures and irrespective of roles, the distributions of decisions reveal multiple peaks: one at Option 1, i.e., full equality, one at 20, i.e., least equality, and in the case of the EF procedure, another one at Option 11, which is one of the two options that are closest to the center.



(a) EF Procedure (b) EQ Procedure

Note: The two panels of the figure show subjects' decisions from 1 to 20 split by the respective relevant roles. The left panel shows the data for the EF Procedure. Role A (shaded) profits from higher options while Role B (light) profits from lower options. Similarly, the right panel shows the data for the EQ Procedure. Role a (shaded) profits from higher options while Role b (light) profits from lower options.

Figure 2: Decisions by Role

In line with Hypothesis 1, differences that depend on subjects' roles are apparent within both procedures. For the EF Procedure, the median of the chosen options by subjects in Role A is 10, while for subjects in Role B it is only 6. Similarly, the average option chosen by those in Role A is 9.81 and only 7.21 for subjects in Role B, a difference of 0.37 standard deviations. These numbers suggest that, indeed, subjects who would themselves profit from high options choose higher options than subjects who would personally profit from low options.

Table 3 analyses the data in a regression framework, regressing subjects' choices on their roles. Its first two column estimate the preregistered regressions equations, i.e., Equation 4. For the EF Procedure, Column 1 shows that the above-mentioned difference in means of 2.6 is statistically significant at any conventional level ($p < 0.001$; two-sided). The same qualitative result of higher choices by subjects in Role A is also confirmed by a non-parametric Mann–Whitney U test ($p < 0.001$; two-sided). The results for the EQ Procedure are qualitatively the same and in quantitative terms slightly stronger. Here, the median option chosen by subjects in Role a is 5, while it is 1 for subjects in

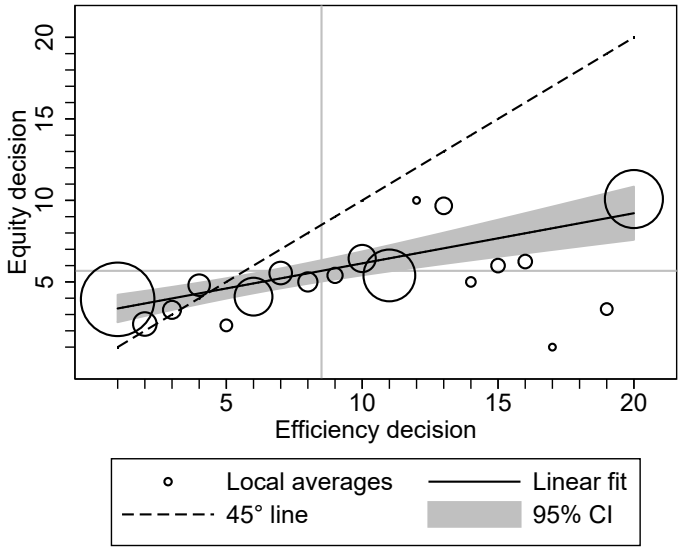
Role b. The means are 7.25 and 4.11, respectively. The difference between the latter values corresponds to 0.47 standard deviations and is thus even larger than the one observed for the EF Procedure. Column 2 of Table 3 shows that this difference is significant ($p < 0.001$; two-sided), and the result is again confirmed by a Mann–Whitney U test ($p < 0.001$; two-sided). Together, the results from both procedures provide clear support for Hypothesis 1, namely for egocentric norm adoption: subjects tend to follow fairness evaluations such that, if the same standards were adopted by everybody, they would personally profit—and their respective group partners would lose.

Table 3: Decisions

Dependent variable	*Decision for succeeding group*			
Model	*OLS*		*SUR*	
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Role A	2.602***		2.609***	1.304*
	(0.727)		(0.722)	(0.674)
Role a		3.140***	1.224*	3.147***
		(0.680)	(0.722)	(0.674)
Constant	7.209***	4.108***	6.593***	3.456***
	(0.498)	(0.428)	(0.625)	(0.584)
Observations	372	372	372	
R^2	0.033	0.055	0.041	0.064

Note: Heteroscedasticity-consistent standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

While the analyses in Columns 1 and 2 have considered subjects' choices for the two procedures in isolation, it is natural to think that they are related. In particular, the fairness tradeoffs in the two procedures both involve the criterion of equality, once weighted against efficiency (EF Procedure), and the other time against the equity principle (EQ Procedure). If a subject puts a strong emphasis on equality, this should show in low choices for both procedures, and a subject who does not consider equality to be very important would be expected to make high choices for both procedures. This means that one would expect choices for the two procedures to be correlated among subjects. Figure 3 displays the empirical relationship between the two decisions that subjects are making. It groups players according to their choices for the EF Procedure, displayed on the horizontal axis, and shows the respective average of their choices for the EQ Procedure on the vertical axis. The sizes of circles correspond to the relative number of subjects. A clear positive trend can be observed in these conditional means. The positive relationship is confirmed by the upward-sloping regression line, which is based on the disaggregated data and corresponds to a correlation of 0.33 ($p < 0.001$, two-sided). This correlation cannot have



Note: The figure groups subjects by their decisions for the EF Procedure. For the subjects belonging to each respective option on the horizontal axis, it plots their average decision for the EQ Procedure on the vertical axis. The sizes of circles correspond to the respective numbers of subjects. The 45-degree line is dotted, and the vertical and horizontal gray lines indicate the averages of decisions for the EF Procedure and the EQ Procedure, respectively. The solid black line represents the linear fit from an OLS regression and the shaded area around it to the corresponding 95% confidence interval based on heteroscedasticity-consistent standard errors.

Figure 3: Relationship Between the Two Decisions

been induced by roles because they are independent. Further inspection also shows that subjects' choices do not seem to be significantly driven by concerns about ex-ante equality. Note that the implications of high choices in terms of procedural fairness are quite different between affected groups in which one of the players has Roles A and a (*parallel* groups) and other (*crossed*) groups. For crossed groups, high choices offset each other from an ex-ante perspective, while they cumulate for parallel groups. Thus, one could expect that decisions over players in crossed groups, which are made subjects in parallel groups, should generally be higher and more strongly positively correlated. Empirically, however, there is no indication for the existence of any of these two effects. The distributions of decisions according to both procedures are not significantly different between the types of groups ($p = 0.28$ for the EF Procedure and $p = 1.00$ for the EQ procedure; Kolmogorov–Smirnov test, two-sided), the positive correlation between decisions holds within both the crossed and the parallel group types ($p < 0.001$ and $p < 0.01$, respectively), and there is no evidence for a difference in the magnitudes ($p = 0.32$).

Given that subjects seem to be consistent in how much weight they attribute to the equality norm in their two decisions, it is useful to consider both procedures jointly for two reasons. First, Hypothesis 2 predicts spillover effects of roles across procedures because the biased weight attached to equality should affect both decisions. Second, the residuals in Columns 1 and 2 are positively correlated because, e.g., a subject who—conditional on her

role—chooses a high option according to the EF Procedure also tends to choose a high option according to the EQ Procedure. Taking into account this correlation structure by estimating equations jointly in a *seemingly unrelated regressions* (SUR) framework (Zellner, 1962) can improve the precision of estimates. Columns 3 and 4 of Table 3 again consider the EF Procedure and the EQ Procedure, respectively, but they are jointly estimated using SUR and each include the effects of both roles. Since the roles in the two procedures are independent, the coefficients of Role A for the EF Procedure and Role a for the EQ Procedure remain virtually unchanged compared to Columns 1 and 2, and the standard errors are slightly smaller.[18] The two other coefficients capture the spillover effects. As predicted by Hypothesis 2, both point estimates are positive and consistent with the interpretation that changes in subjects' fairness judgments induced by roles carry over to the respective other procedure in a manner that is similar to preexisting differences between different individuals.[19] Individually, both spillover coefficients are weakly statistically significant ($p < 0.1$), and they are jointly significant at the five percent level ($p = 0.04$).

This was the starting point for developing Hypothesis 2. Given that roles lead to egocentric biases *within* procedures and subjects apply (at least somewhat) consistent fairness judgments to both procedures, changes in subjects' fairness judgments induced by roles should carry over to the respective other procedure in a manner that is similar to preexisting differences between different individuals.Hypothesis 2 thus predicts spillovers of roles *across* procedures. Columns 3 and 4 of Table 3 regress decisions for both procedures on the subjects' respective irrelevant groups. As predicted, both coefficients are positive.

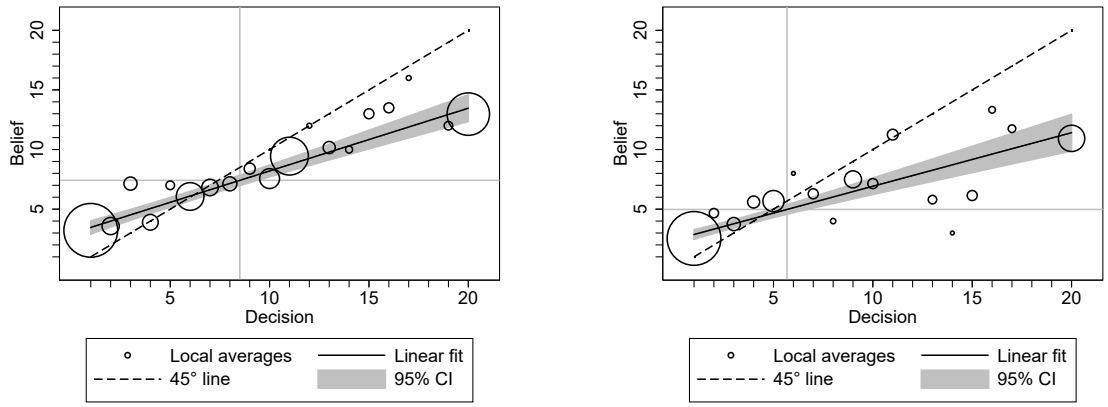
4.2 Beliefs

An important question is whether the egocentric behavior of subjects is conscious or, as assumed in the formal framework presented in Section 3, unconscious, i.e., whether and to which extent subjects realize that their decisions deviate from their true fairness convictions. To address this point, we next analyze subjects' beliefs about average choices made by others.

A first important observation can be made from Figure 4. Both of its panels are similarly constructed as the Figure 3 in Section 4.1 above, but plotting average beliefs for subjects depending on their decisions. Figure 4a on the left shows the relationship for the

[18] The point estimates slightly differ because, as mentioned earlier in Footnote 14, four subjects did not complete the experiment, and roles are therefore not precisely independent anymore. The slight empirical correlations between roles are—of course—random, and the implications for estimates minimal.

[19] Multiplying the effect of Role A with the slope of the regression line in Figure 3 (0.31), one would expect an effect of Role A for the EQ Procedure of 0.80. Similarly, the respective prediction for the effect of Role a for Procedure EF would be 1.08. The observed values in Columns 3 and 4 of Table 3 are comparable to these predictions and even slightly larger.



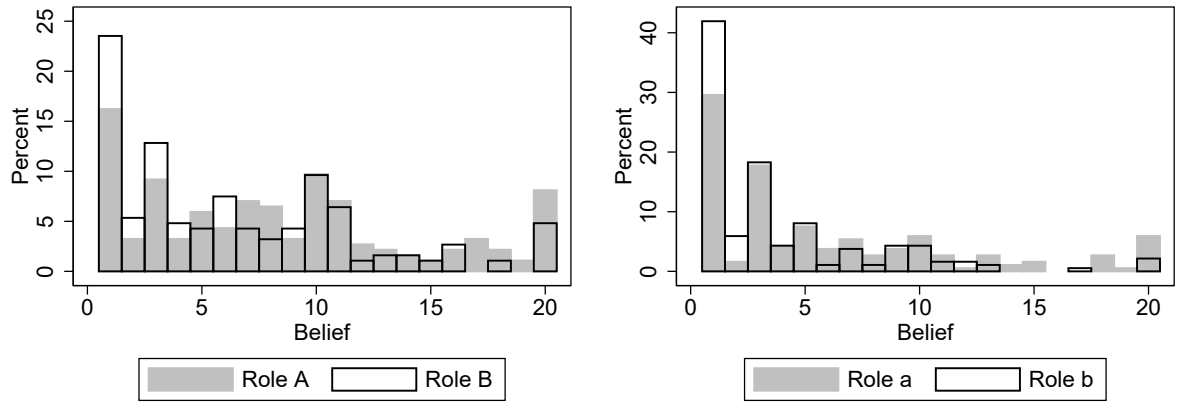
(a) EF Procedure (b) EQ Procedure

Note: The two panels of the figure group subjects by their decisions for the EF Procedure and the EQ Procedure, respectively. For the subjects belonging to each respective option on the horizontal axis, it plots their average beliefs about others' decisions for the same procedure on the vertical axis. The sizes of circles correspond to the respective numbers of subjects. The 45-degree lines are dotted, and the vertical and horizontal gray lines indicate the averages of decisions and beliefs, respectively. The solid black lines represent the linear fits from OLS regressions and the shaded areas around them to the corresponding 95% confidence intervals based on heteroscedasticity-consistent standard errors.

Figure 4: Beliefs vs. Decisions

EF Procedure and Figure 4b on the right does the same for the EQ Procedure. Average decisions and beliefs are marked by the light lines that run vertically and horizontally, respectively. In both panels, they intersect below the 45-degree lines, meaning that average beliefs are lower than average decisions for both the EF Procedure and the EQ Procedure ($p < 0.001$ and $p = 0.01$, respectively), which means that subjects typically expect others to assign a higher relative weight to equality than they do themselves. Moreover, in both instances, there exists a clear positive correlation between choices and beliefs ($p < 0.001$), with slopes of 0.53 for the EF Procedure and 0.45 for the EQ Procedure. It is unlikely that these relationships stem from cognitive dissonance reduction as, e.g., in Konow (2000). That is, first, because beliefs were elicited using an incentivized procedure (see Section 2), i.e., it would have been costly for subjects to bias their beliefs. And, second, because there was no selfish motive that subjects could wish to conceal. Thus, the positive associations between decisions and beliefs are better interpreted as reflecting the well-established false consensus effect (Ross, Greene, and House, 1977): people have a fundamental disposition that others' convictions are more similar to their own than they actually are.

The latter was the premise under which Hypothesis 3 was derived, which states that egocentric biases should also be present in beliefs. As Figure 5 shows, this is the case. The two panels replicate those of Figure 2, replacing choices with beliefs about others' choices. For both procedures, the distributions of beliefs feature more weight at values around the center than it is observed for choices. This is consistent with individuals forming beliefs by using their own choices as anchors and then adjusting them towards more moderate



(a) EF Procedure (b) EQ Procedure

Note: The two panels of the figure show subjects' beliefs about others' average decisions from 1 to 20 split by the respective relevant roles. The left panel shows the data for the EF Procedure. Role A (shaded) would profit from higher options while Role B (light) would profit from lower options. Similarly, the right panel shows the data for the EQ Procedure. Role a (shaded) would profit from higher options while Role b (light) would profit from lower options.

Figure 5: Beliefs by Roles

values. With respect to role differences, however, the pattern is the same as for choices: Subjects in roles A and a (shaded) expect higher choices than those in roles B or b (light), respectively. The visual impression is confirmed by the corresponding regression coefficients presented in Columns 1 and 2 of Table 4. Both the effect of Role A for the EF Procedure and the one of Role a for the EQ Procedure are positive and statistically significant ($p < 0.01$ and $p < 0.001$). In terms of standard deviations, the effects amount to 0.32 for the EF Procedure 0.42 for the EQ Procedure, which are relative sizes that are very close to those found for decisions.[20]

For decisions, Figure 3 had established a strong positive correlation of 0.33 between the two procedures. Figure 6 shows the corresponding relationship for beliefs. While there exists a clear positive relationship as well ($p < 0.001$, two-sided), the correlation of 0.18 is much weaker than the one for decisions. In light of this finding, it is not surprising that spillover effects in beliefs, which are reported in Columns 3 and 4 of Table 4, are small and not statistically significant. However, the point estimates are still positive, and the fact that they cannot be distinguished from zero is, of course, no evidence against their existence. Rather, it points towards insufficient statistical power for identifying this subtle effect.

Overall, the analysis of beliefs yields patterns that are very similar to those for decisions, and effect sizes are about as large as if the exogenously induced treatment effects

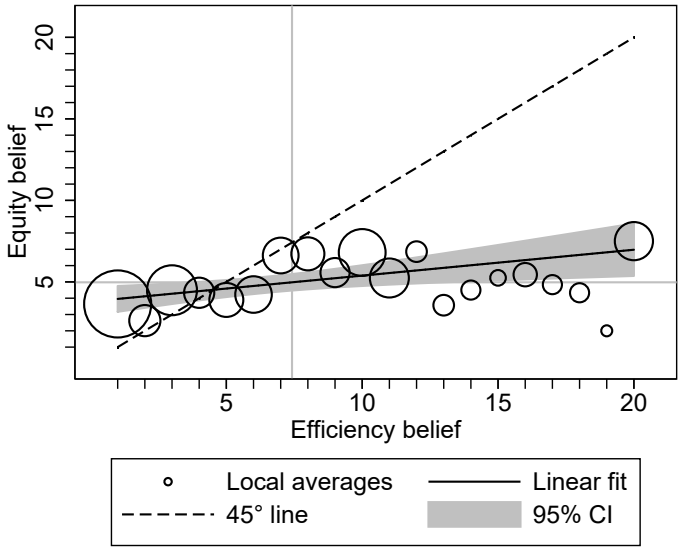
[20]The expected sizes of coefficients as the result of treatment effects on decisions and relationship between decision and beliefs can be calculated following the same logic as for spillover in Footnote 19 by multiplying the respective effects found for decisions with the slopes from Figure 4. The predicted effects for beliefs amount to 1.37 for the EF Procedure and to 1.41 for the EQ Procedure. The actually observed effects are similar to these predictions and—just as observed for spillovers in Table 3—even a bit larger.

Table 4: Predictions

Dependent variable	*Belief about others' average decisions*			
Model	*OLS*		*SUR*	
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Role A	1.849***		1.852***	0.231
	(0.585)		(0.582)	(0.508)
Role a		2.102***	0.639	2.103***
		(0.510)	(0.582)	(0.508)
Constant	6.497***	3.925***	6.176***	3.809***
	(0.390)	(0.299)	(0.504)	(0.440)
Observations	372	372	372	
R^2	0.026	0.044	0.029	0.045

Note: Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

had been due to preexisting differences between subjects (which is, of course, rules out by random assignment to roles). Thus, the results suggest that the treatment effects arise (perhaps entirely) unconsciously and that subjects are largely unaware of the egocentrism that is present in their fairness judgments.



Note: The figure groups subjects by their beliefs about others' average decisions for the EF Procedure. For the subjects belonging to each respective option on the horizontal axis, it plots their average belief about others' decisions for the EQ Procedure on the vertical axis. The sizes of circles correspond to the respective numbers of subjects. The 45-degree line is dotted, and the vertical and horizontal gray lines indicate the averages of decisions for the EF Procedure and the EQ Procedure, respectively. The solid black line represents the linear fit from an OLS regression and the shaded area around it to the corresponding 95% confidence interval based on heteroscedasticity-consistent standard errors.

Figure 6: Relationship Between the Two Predictions

4.3 Further Observations

By the design of the experiment, roles do not induce differential proximity between players in adjoining groups, since roles are crossed. Independently of roles, however, the design deliberately induces *nominal* groups by referring to players in each group as X and Y. Thus, participants decide over allocations between one player with the same name as themselves and one with a different name. In this dimension, the experiment mimics research on discrimination between *minimal groups* in social psychology (Tajfel, Billig, and Bundy, 1971; Billig and Tajfel, 1973) and economics (Chen and Li, 2009). If subjects favored their nominal in-group, they should choose a high option for EF Procedure if the receiving player sharing their name is in Role A, and a high option for the EQ procedure if the player is in Role a. Table B.2 shows the corresponding results for decisions (Columns 1 and 2) and beliefs (Columns 3 and 4). In line with the findings in the literature, subjects exhibit significant nominal in-group bias in both procedures ($p < 0.01$ for the EF and $p < 0.001$ for the EQ Procedure). The effect sizes are smaller than the ones estimated for roles, although the differences are not significant. To a smaller extent, corresponding effects are also found for beliefs. Both estimated coefficients are positive, and the effect for the EQ procedure is significant ($p < 0.001$). Since names were determined independently of names, the estimated effects of roles and names are virtually unaffected by including the respective regressors jointly, as is shown in Table B.3 (see the appendix).

A potential concern with regard to almost all experiments involving human subjects is that effects might be driven by experimenter demand, i.e., subjects being able to guess the research hypothesis and trying to conform with the expectations of the experimenter. This experiment was designed such as to mitigate such concerns as much as possible. An important design property is that treatment effects are identified between-subjects, as opposed to within-subjects. This comes at the cost of not being able to calculate individual-specific effects, but subjects are not made aware of the treatment differences or their own counterfactual behavior. In fact, in studying group bias, Chen and Li (2009) rely mainly on a within-subject design and use a between-subject treatment specifically to mitigate experimenter demand effects. As discussed above, this experiment studies (nominal) group bias as well, and this decision was, in part, also made to conceal the purpose of the design. Should subjects have tried to guess the research hypotheses, they might have ended up with the wrong one, or they would have had to balance multiple conflicting motives. Under these conditions, it seems implausible that the observed effects could be so large. For specific "demand treatments," De Quidt, Haushofer, and Roth (2018) find average effects of 0.13 standard deviations. In contrast, the effects observed in this experiment are multiple times as large. The effects can also be observed for beliefs, which were incentivized. Here, subjects would have had to give up their own money. Moreover, the data generally seem well-behaved. For example, Table B.4 (see the appendix) shows that the randomly determined order of procedures matters for effect sizes in a conceivable way: effects on decisions are stronger for the respective procedure that comes first, although not significantly. And lastly, the next section will show that the treatment effects are not mainly driven by a view subjects making extreme decisions but by the bulk of subjects making decisions that seem to be moderately biased.

5 Heterogeneity

After having established the effect of subjects' own roles on decision making for others, this section aims at relating the effect to relevant personal attributes of subjects. In terms of determinants, we consider the role of different aspects of empathy as well as prosociality. Regarding outcomes, we study the relationship with progressivism and political orientation on the left–right spectrum.

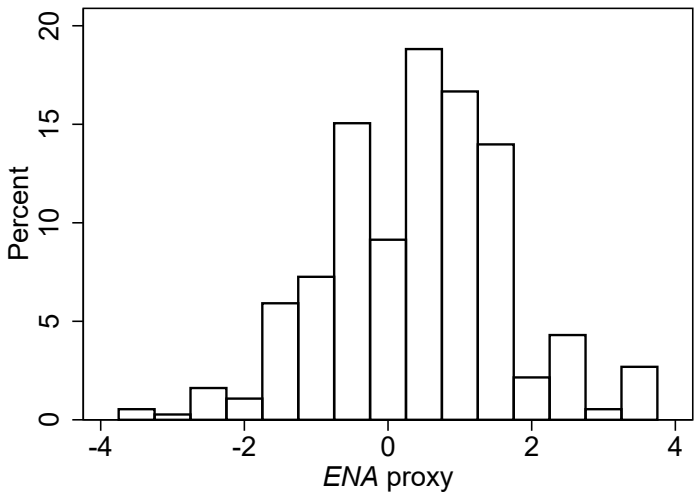
A challenge for studying individual heterogeneity in the display of egocentric norm adoption is that the treatment effects of roles are identified not within but only *between* subjects. We therefore first discuss how the two decisions that any subject is taking can be converted into a single individual-specific proxy of egocentric norm adoption, denoted by *ENA*. We start from the self-evident fact that, if there was no treatment effect, it would make no difference for a given subject's decisions to which roles she has been assigned. The average choices conditional on roles would thus coincide the unconditional

average answers. A measure for how much a particular choice contributes to the treatment effect is thus given by how much it deviates from the unconditional average choice in the direction that favors the subject's relevant role. For every decision, we therefore calculate the deviation from the average of choices for the respective procedure. For better comparability across procedures, we further divide the differences by the respective standard deviation, i.e, we transform subjects' choices into z-scores denoted by by $z_{i,g}^{EF}$ and $z_{i,g}^{EQ}$ for the EF Procedure and the EQ Procedure. The *ENA* proxy is then constructed by adding the respective z-score if a subject has the relevant roles A or a and by subtracting it if the role is B or b. Using the indicator function 1_A for whether subject i in group g's role for the EF Procedure, $r_{i,g}^{EF}$, is A and the analogous indicator function $1_a\left(r_{i,g}^{EQ}\right)$, the *ENA* proxy is thus calculated as follows:

$$ENA_{i,g} \equiv \left[1_A\left(r_{i,g}^{EF}\right) - 1_B\left(r_{i,g}^{EF}\right)\right] z_{i,g}^{EF} + \left[1_a\left(r_{i,g}^{EQ}\right) - 1_b\left(r_{i,g}^{EQ}\right)\right] z_{i,g}^{EQ} \tag{5}$$

Deviations that are aligned with a subject's relevant role contribute to higher values of *ENA*, while deviations that are opposed to the relevant role lead to a decrease. A subject who makes average decisions for both procedures receives a value of zero, irrespective of her roles. On the other hand, a subject who, e.g., makes a high decision for the EF Procedure and a low decision for the EQ Procedure receives a large positive value if her roles are (A, b), values closer to zero if her roles are (A, a) or (B, b), and a large negative value of *ENA* if her roles are (B, a). For any individual subject, a high or low value of the *ENA* proxy can, of course, be entirely due to her true fairness convictions, which simply happen to coincide or conflict with the interests of her roles, leading to noise in the proxy. However, the proxy can be informative when many subjects are considered together.

Figure 7 shows the distribution of the *ENA* proxy in the full sample. Its mean value is 0.42, which is by construction equal to the mean of the two treatment effects in terms of standard deviations (0.37 for the EF Procedure and 0.47 for the EQ Procedure; see Section 4.1). In line with the previous findings, this positive average is significantly different from zero ($p < 0.001$, two-sided t-test). The figure shows that the positive average value, i.e., the effects of roles, is not mainly driven by a few subjects at the extremes but importantly also by many subjects who exhibit moderate levels of bias in the direction of their roles' interests. When restricting the sample to, e.g., only those 262 out of 372 subjects for whom the value of the *ENA* proxy lies in the interval $[-1.5, 1.5]$, the average value is still significantly positive ($p < 0.001$, two-sided t-test).



Note: The figure shows the distribution of the *ENA* proxy calculated according to Equation 5 for the full sample, using a bin width of 0.5.

Figure 7: Distribution of the *ENA* proxy

5.1 Survey Measures

After the main experiment, subjects completed a number of questionnaires that were selected to measure potentially relevant personal characteristics. The elicited groups of characteristics are introduced below.

Empathy From a conceptual perspective, empathy is the personality trait that is most directly relevant for the effect under study. To measure empathy, we use the well-established Interpersonal Reactivity Index (IRI) developed by Davis (1980), which consists of four subscales. The first, *perspective-taking*, should be of particular importance for non-egocentric behavior (Davis, 1983). It is measured with questions such as "I believe that there are two sides to every question and try to look at them both" (p. 11). Higher scores thus indicate that people typically make an effort to "put themselves in others' shoes", i.e., that they should tend to abstract from their roles in the experiment. Second, *fantasy* measures people's tendency to identify with fictitious characters, e.g., in books or movies. Third, *empathic concern* captures the extent to which people feel for others who are in need. And fourth, *personal distress* addresses whether people feel anxious when they witness others' suffering. While the first three aspects usually have a positive connotation, the last one is usually seen as standing in the way of compassionate behavior because it makes people turn away from others' problems. In the experiment, therefore, personal distress could mean that people in advantaged positions tend to avoid thinking about those others who are disadvantaged and, therefore, might be more prone to taking egocentric decisions.

Prosociality The experiment presented in this paper has been designed to show a bias that speaks of egocentrism. In contrast, egoism has been muted in the experiment due to the absence of selfish incentives. To study the role of prosociality empirically, however, the questionnaires included the qualitative items from the Preference Survey Module (Falk et al., 2016) for altruism, positive reciprocity, and trust, which are subsumed under *prosociality* in Falk et al. (2018).

Values Conceptually, the personality traits of empathy and prosociality are potential *determinants* of egocentric norm adoption. Since the biases that are observed in the experiment seem to arise unconsciously, they are unlikely to be the result of people's values. Instead, we will consider the latter as potential *outcomes*. A leading approach in modern moral philosophy to understand how moral values vary across the political spectrum is Moral Foundations Theory (MFT) (Haidt and Joseph, 2004; Haidt and Graham, 2007; Graham, Haidt, and Nosek, 2009), which traces (cultural) differences in ethical judgments to the respective weights attached to five distinct dimensions of moral intuitions: *harm/care*, i.e., being compassionate with those in need; *fairness/reciprocity*; *ingroup/loyalty*; *authority/respect*; and *purity/sanctity*. We included the 30-item Moral Foundations Questionnaire (MFQ) that was created by a group of researchers around the developers of MFT.[21] As suggested by the developers, we aggregate the five subscales into a single measure of *progressivism*.[22]

$$\begin{aligned} progressivism = &(harm/care + fairness/reciprocity) \div 2 \\ &- (ingroup/loyalty + authority/respect + purity/sanctity) \div 3 \end{aligned}$$

We also include a simple question about people's political attitude on scale from *left* to *right* (European Social Survey, 2014). The variables *progressivism* and *political attitude* turn out to be highly correlated in the expected direction ($r = -0.51$, $p < 0.001$).

Personality Controls The further qualitative preference items by Falk et al. (2016) for risk preferences, time preferences, and negative reciprocity are included as control variables. Moreover, the questionnaires included the Big Five personality inventory, which is probably the most widely used framework to study people's personalities. Specifically, we use a translation of the 15-item BFI-S scale developed by Gerlitz and Schupp (2005). The Big Five traits are: *openness*, capturing interest in new experiences; *conscientiousness*, encompassing whether a person is determined and organized; *extraversion*, i.e, how much people like to engage with others; *agreeableness*, measuring altruistic motivation and

[21] The questionnaire is is publicly available on the web (https://moralfoundations.org/questionnaires/; retrieved in May 2020).

[22] A very similar measure is used by Enke (2020), who excludes the *purity/sanctity* dimension and focuses on communal vs. universal values in the context of political competition. In my data, these two measures based on the same questionnaire have a correlation of 0.96, i.e., they hardly differ.

cooperative behaviors; and *neuroticism*, referring to emotional instability and anxiety.

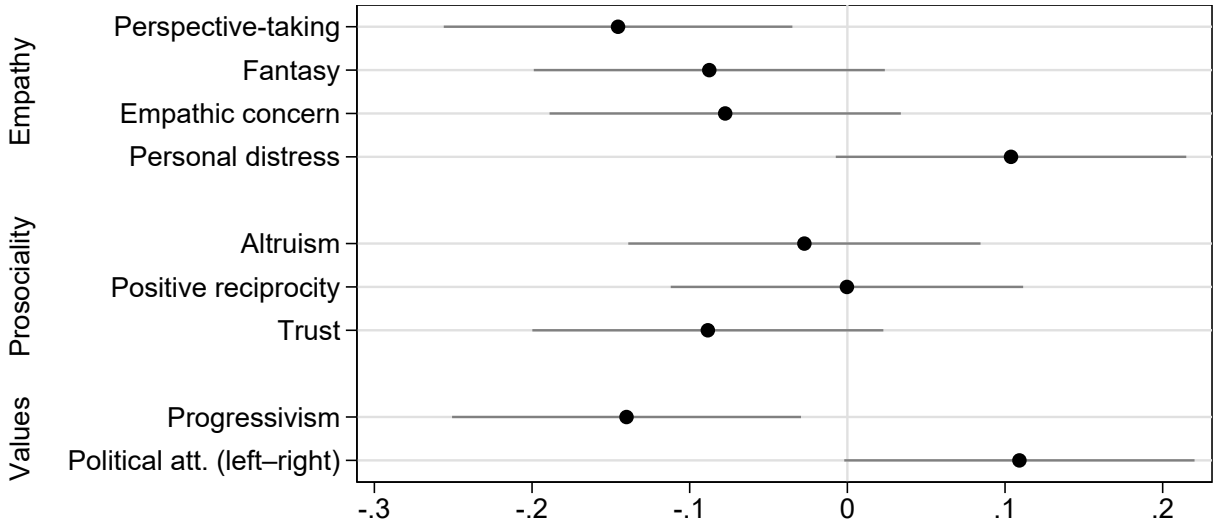
Demographic Controls The elicited sociodemographic controls are subjects' gender (female, male, or diverse), their age, enrollment at a university, and gross monthly income in euros. The latter is converted into log income as $\ln(income)$.

In studying heterogeneity in the treatment effects, we restrict the sample to individuals who seem to have had no major difficulties in understanding the experiment and took answering the survey questions seriously. The experiment included a number of control questions, and subjects could only progress once they had answered them all correctly. It is, however, conceivable that those who initially might not have properly understood the experiment.[23] Therefore, during the experiment, we automatically recorded the number of incorrectly submitted questions. We exclude those subjects above the 90^{th} percentile of the distribution of mistakes. After the questionnaires, we asked subjects about the reliability of their answers. We exclude subjects how gave answers below the tenth percentile. This leaves me with 312 subjects, for whom all previous results from Section 4 replicate. Results for the full sample are provided in Appendix B.

5.2 Observed Heterogeneity

To get an overview of how the characteristics of interest are related to the treatment effects of roles in the experiment, Figure 8 considers the correlations between the *ENA* and the different measures of empathy, prosociality, and values. For the four dimensions of empathy, a pattern arises that is consistent with the theoretical predictions: the three "positive" facets of empathy—perspective-taking, fantasy, and empathic concern—are negatively correlated with egocentric norm adoption, i.e., higher empathy in these regards leads to lower egocentric bias. On the other hand, the "negative" side of empathy—personal distress—leads to a stronger egocentric bias, potentially because subjects avoid thinking about being in a position that would upset them. The correlation with perspective-taking, which is the opposite of egocentrism, is significantly negative ($p = 0.01$, two-sided), and the correlation with personal distress is (weakly) significantly positive ($p = 0.07$, two-sided). The correlations with fantasy and empathic concern are not statistically significant ($p > 0.1$, two-sided). For all of the prosociality measures, no significant correlations are observed ($p > 0.1$). In particular, the correlation between the *ENA* proxy and altruism is close to zero, consistent with the irrelevance of selfishness. For moral values, we find a negative correlation with the ENA proxy for progressivism ($p = 0.01$), constructed using the MFQ. People holding liberal values thus seem to exhibit less of the egocentric bias than conservatives. In light of this finding, it perhaps not surprising to find that people

[23] This could be due to a lack of effort. Other problems, such as language barriers, also seem plausible, given that the experiment had to be conducted online.



Note: The figure shows the Pearson correlation coefficient for the ENA proxy introduced in Equation 5 and the respective survey measure. Gray bars indicate 95% confidence intervals. The analysis excludes subjects above the 90^{th} percentile in the distribution of mistakes in the control questions and those whose self-reported reliability regarding the survey answers lies below the 10^{th} percentile, leaving 312 subjects.

Figure 8: Correlations with the *ENA* Proxy

leaning to the political right show a larger bias than those leaning to the left ($p = 0.05$).[24] While the significance for the correlation with progressivism only holds in the restricted sample that is used in this section, the qualitative results for empathy and prosociality are the same for the full sample (see Figure B.1 in the appendix) Interestingly, among the demographic control variables (age, female vs. male, student, and log income; not shown), no significant correlations with the *ENA* proxy are observed.

The analysis of heterogeneity is deliberately descriptive and does not aim at making causal claims. However, because many of the variables considered above are correlated, it would be interesting to see if the observed correlations with the potential determinants, i.e., with the different facts of empathy and prosociality, merely reflect different symptoms of maybe just a single underlying relationship or whether they also hold conditionally on each other. We therefore employ a multivariate regression framework. The first columns shows the results of regressing the *ENA* dummy simultaneously on all four facets of empathy, as in all other columns controlling for other personality characteristics (see Section 5.1 above) and showing standardized coefficients. The estimates confirm the insights from Figure 8. Strong perspective-taking is associated with a smaller egocentric bias in decision making ($p = 0.02$). The coefficients for fantasy and empathic concern point towards the same direction but remain insignificant ($p > 0.1$), while the positive association between the *ENA* proxy and personal distress is now significant ($p < 0.01$). Column 2 considers

[24]Progressivism and political attitude are strongly correlated in my data ($r - 0.51$ in the full and $r = -0.49$ in the restricted sample; both $p < 0.001$, two-sided). The correlations of progressivism and universalism in the data is $r = 0.96$ ($p < 0.001$). Results for the two variables are very similar but a bit stronger for progressivism (which uses more items).

Table 5: Heterogeneity

Dependent variable	*ENA proxy*			
	(1)	(2)	(3)	(4)
Perspective-taking	-0.179**		-0.176**	-0.185**
	(0.0748)		(0.0754)	(0.0754)
Fantasy	-0.0755		-0.0999	-0.0886
	(0.0728)		(0.0760)	(0.0786)
Empathic concern	-0.0303		-0.00779	-0.0109
	(0.0718)		(0.0783)	(0.0792)
Personal distress	0.236***		0.248***	0.261***
	(0.0718)		(0.0712)	(0.0719)
Altruism		-0.0442	-0.00313	-0.00408
		(0.0635)	(0.0713)	(0.0718)
Positive reciprocity		0.0111	0.0400	0.0473
		(0.0608)	(0.0601)	(0.0611)
Trust		-0.0956	-0.111*	-0.104*
		(0.0620)	(0.0606)	(0.0618)
Personality controls	Yes	Yes	Yes	Yes
Demographic controls	No	No	No	Yes
Observations	312	312	312	312
R^2	0.108	0.054	0.120	0.130

Note: The table reports standardized coefficients. Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. The analysis excludes subjects above the 90^{th} percentile in the distribution of mistakes in the control questions and those whose self-reported reliability regarding the survey answers lies below the 10^{th} percentile, leaving 312 subjects. The *personality controls* are risk and time preferences along with the Big Five traits. The *demographic controls* are gender (categories: female, male, and diverse), age and squared age, a dummy for being enrolled at university, and the log of monthly gross income in euros, calculated as $\ln(income + 1)$.

the three measures of prosociality and does not find significant coefficients for either of them. Column 3 considers the variables pertaining to empathy and prosociality jointly. The estimates remain stable and the only minor qualitative change is that the negative coefficient for trust becomes (weakly) statistically significant ($p = 0.07$). The final column adds demographic controls for capturing further influences that might not be accounted for by the personality controls. However, the effect on estimates is negligible. All of the qualitative results from the Table replicate also for the full sample of all 372 participants (see Table B.5 in the appendix).

Overall, the main result from the analysis of heterogeneity in the treatment effect of roles is that empathy contributes most to explaining differences between individuals. The empirical associations between egocentric norm adoption and empathy are consistent with the theory that underlies the measures that were used: perspective-taking is negatively

associated with the strength of the egocentric bias that is observed in the experiment, while personal distress is associated with an increased bias.

6 Discussion

This paper has provided experimental evidence for the phenomenon of egocentric norm adoption, whereby people act according to norms from which they would personally profit if they were adopted by others. The key property of the experiment was that people's own decisions were in no way relevant for their own payoffs but that subjects depended on others' choices in the same decision contexts. Points were distributed between subjects within groups of two players according to two procedures. One of these implied a tradeoff between equality and efficiency, and the other involved the norms of equality and equity. Groups were numbered consecutively, and the players of each group decided over the subjects in the respective succeeding groups. Across adjoining groups, the roles of players (in terms of favoring norms) were crossed, such that players shared exactly one role with each subject over which they decided. An egocentric bias was found for both procedures, and corresponding biases of similar size were also found in subjects' beliefs about others' behavior. The analysis of heterogeneity provides additional support in favor of egocentrism as the key driving force behind the treatment effects of roles: the bias is largest among subjects who report weak perspective-taking.

This last point also speaks against an interpretation of the results in the sense of subjects confusing *diagnostic* with *causal* contingencies (Quattrone and Tversky, 1984; Shafir and Tversky, 1992), whereby subjects would try to "induce" a desired behavior by others with their own actions. Similarly, "wishful thinking" (see, e.g., Mijovic-Prelec and Prelec, 2010; Engelmann et al., 2019) would not predict the observed correlations with the empathy subscales. Nonetheless, future research could adapt the experimental design of this paper with the modification of putting some subjects into a position where their own payoffs are determined by the computer, i.e., a non-human random device. Less biased behavior in such a treatment would indicate that subjects' decisions partly also reflect a concern with the choices of others.

This paper on egocentric norm adoption is closely related to a paper by Hofmeier and Neuber (2019) that provides experimental evidence for a slightly more specific phenomenon termed *imperfect empathy*. In the experiment that the paper presents, subjects are sequentially paying for themselves and others for avoiding disgust. First, their willingness to pay (WTP) is elicited for *not* having to eight different food items containing dried insects. Then, they are in the role of a *sender* who has to report her WTP for a *receiver* not having to eat one particular of the eight items. This is done for a succession of eight receivers in total, each time for a different item. Each sender also serves as a receiver, and, crucially, senders are informed about the full vectors of receivers' WTPs

for themselves. The main result of the paper is that WTPs for others are high if both the sender's and the receiver's personal WTPs are high and low otherwise. The importance of senders' WTPs is even found *within subjects*, i.e., a given sender behaves more altruistically when an item is concerned that she dreads rather than one for which her own WTP is relatively low. The experiment on imperfect empathy differs from the one in this paper in that tradeoffs are not made between two others but between *oneself* and another subject. Moreover, preferences are not exogenously induced but, instead, preexisting heterogeneity in subjects' preferences is being used. However, the results of both experiments are instances of egocentric norm adoption: people act like they would want others to act, i.e., they apply norms from which they would personally profit if they were followed by others, and their empathy is triggered if they would value help themselves. The findings from both experiments also caution against the equivalent use of elicitation procedures for social preferences with our without role uncertainty. In line with egocentric norm adoption, Iriberri and Rey-Biel (2011) and Zhan, Eckel, and Grossman (2020) find increased pro-sociality in (modified) dictator games when it is ex-ante uncertain whether a given subject will be paid according to one of her own decisions as a dictator or according to a decision by another subject as the receiver.

The idea of "acting like one would want others to act" is related to the concept of *rule-utilitarianism* advocated as a normative principle by Harsanyi, 1977, whereby "an individual act should be considered to be morally right if it conforms to the correct moral rule applying to this type of situation – regardless of whether it is the act that will or will not yield the highest possible social utility on this particular occasion" (p. 32). In particular, Harsanyi applies the logic of rule-utilitarianism to voting contexts. He shows that, if people were following rule-utilitarianism, this would to some extent resolve the *paradox of voting*, which describes the seemingly irrational behavior of people who incur the costs of voting in large elections (e.g., in terms of time) while almost certainly not being pivotal for the outcome (Downs, 1957). Rule-utilitarianism is an abstract normative concept that most potential voters are most likely not familiar with, but egocentric norm adoption could explain why many people intuitively conform with it: like the subjects in the experiment by Hofmeier and Neuber (2019), they incur costs because they would like others to do the same. In the examples discussed by Harsanyi (1977), certain thresholds for the number of votes must be achieved for the socially optimal option to be implemented, e.g., because a fixed number of votes is cast in favor of the respective alternative option that is socially suboptimal. Harsanyi does not discuss how these votes come about. Under the label of *ethical voting*, some contributions have made suggestions for positive theories that resolve the paradox of voting. Feddersen and Sandroni (2006a, 2006b) and Coate and Conlin (2004) develop closely related models of voting over two alternative options. Both approaches assume *ethical* voters who follow rules that they would want to be followed by everybody *who favors the same option* as they do themselves, taking the behavior of

non-ethical voters and ethical voters who favor the opposite option as given.[25] However, one might still be puzzled why people who behave ethically in terms of incurring the (useless) costs of voting should disagree on what is the optimal policy. Egocentric norm adoption offers an explanation: people consider options as fair from which they would personally profit, i.e., the selfish option subjectively is perceived as the one that is ethical. The same underlying behavioral phenomenon can thus account for prosocial behavior in terms of turning out to vote and for selfish behavior with respect to supported policies.

The models by Feddersen and Sandroni (2006a, 2006b) and Coate and Conlin (2004) both feature heterogeneous costs of voting. Because ethical voters aim at maximizing the utility of some group, they adopt rules or strategies which prescribe that ethical voters who share their preference should vote if and only if their costs of voting do not exceed a certain threshold value, because winning by an excessive margin would be wasteful. Thus, in the above models, heterogeneity, in fact, enables coordination between voters who favor the same option. This model implication is, however, at odds with evidence from experiments on public goods games that feature heterogeneity. Here, the typical finding is that heterogeneity reduces efficiency (see Fischbacher, Schudy, and Teyssier, 2014 and references therein), and Kube et al. (2015) find that heterogeneity also makes it more difficult for subjects to agree on efficiency-enhancing institutions, i.e., sets of mandatory rules. Contrary to rule-utilitarianism, egocentric norm adoption is in line with these results since it implies that people will opt for sets of rules in their own favor. Incorporating egocentric rule-following into models of voting and testing the resulting predictions would be an interesting subject of future research.

Beyond voting, many other real-world phenomena can be understood more clearly when considering the egocentric nature of norm adoption. Arguably the most important collective action problem of our time is the fight against global warming, i.e., in particular the need to reduce global carbon dioxide emissions. It is true for all countries in the world that unilateral action is pointless from a strict (act-) utilitarian perspective, as costs are high and private returns (for a given country) are low. This applies to China and the United States, which account for 29.7% and 13.9% of global emissions in 2019, respectively (Crippa et al., 2019), but even more so for, e.g., the Marshall Islands, a small country in the Pacific Ocean that is part of Micronesia. Nonetheless, the country that is endangered by rising sea levels has announced a plan of reducing carbon dioxide emissions to zero by 2050 (Malo, 2018). It is conceivable that behind this step mainly stands the wish for other countries to do the same, thereby hopefully securing the islands' future existence. In this context, the egocentric adoption of norms is also closely linked to setting an example (cf. Gächter et al., 2012; Gächter, Nosenzo, and Sefton, 2013). Indeed, Bicchieri

[25]The two models differ in the objectives that individuals pursue: in the model by Feddersen and Sandroni (2006a), ethical voters maximize the utility of all people, while Coate and Conlin (2004) assume that they maximize only the utility of those people who share their own preferences, i.e., of those who are in their group.

et al. (2020) show that people's propensity to comply with norms is eroded when they observe others breaching the norm, and heightened when others obey the norm. This finding is also suggestive that acting upon norms from which one would want others to follow is particular from an evolutionary perspective: if one's own behavior is observable, it does indeed make it more likely that others will follow the norm. Egocentric norms adoption might, therefore, be an evolutionary selected, automatic way of *endorsing* norms.

In 2020, people around the world have been able to watch new norms being established in real-time in the course of the COVID-19 pandemic. Face masks have become ubiquitous, although they are not protecting the wearer herself but instead other people of whom many are strangers. In light of the findings of this paper, the relative ease with which wearing face masks has been established as a new social norm is probably due to the fact that almost everybody (at least almost all informed individuals) would want others to wear them. In terms of communication, this means that stressing heterogeneous levels of vulnerability for certain subgroups of the population might be counterproductive, i.e., undue emphasis on the fact that young people are in less danger of dying from COVID-19 might hamper sustaining social norms which are crucial for containing the pandemic.

References

Akerlof, George A., and William T. Dickens. 1982. "The Economic Consequences of Cognitive Dissonance". *American Economic Review* 72 (3): 307–319.

Alesina, Alberto, Paola Giuliano, and Nathan Nunn. 2013. "On the Origins of Gender Roles: Womenand the Plough". *The Quarterly Journal of Economics* 128 (2): 469–530.

Alger, Ingela, and Jörgen W. Weibull. 2019. "Evolutionary Models of Preference Formation". *Annual Review of Economics* 11:329–354.

— . 2013. "Homo Moralis—Preference Evolution Under Incomplete Information and Assortative Matching". *Econometrica* 81 (6): 2269–2302.

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving". *The Economic Journal* 100 (401): 464–477.

Babcock, Linda, and George Loewenstein. 1997. "Explaining Bargaining Impasse: The Role of Self-Serving Biases". *Journal of Economic Perspectives* 11 (1): 109–126.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased Judgments of Fairness in Bargaining". *American Economic Review* 85 (5): 1337–1343.

Barron, Kai, Robert Stüber, and Roel van Veldhuizen. 2019. *Motivated motive selection in the lying-dictator game.* Discussion Paper SP II 2019–303. Berlin, Germany: Wissenschaftszentrum Berlin für Sozialforschung (WZB).

Becker, Gary S. 1974. "A Theory of Social Interactions". *Journal of Political Economy* 82 (6): 1063–1093.

— . 1976. "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology". *Journal of Economic Literature* 14 (3): 817–826.

Bénabou, Roland, and Jean Tirole. 2016. "Mindful Economics: The Production, Consumption, and Value of Beliefs". *Journal of Economic Perspectives* 30 (3): 141–164.

Bergstrom, Theodore C. 1995. "On the Evolution of Altruistic Ethical Rules for Siblings". *America Economic Review* 85 (1): 58–81.

Bicchieri, Cristina, and Alex K. Chavez. 2013. "Norm Manipulation, Norm Evasion: Experimental Evidence". *Economics and Philosophy* 29 (2): 175–198.

Bicchieri, Cristina, Eugen Dimant, Simon Gächter, and Daniele Nosenzo. 2020. *Observability, Social Proximity, and the Erosion of Norm Compliance*. CESifo Working Paper 8212. Munich, Germany: Munich Society for the Promotion of Economic Research – CESifo.

Billig, Michael, and Henri Tajfel. 1973. "Social categorization and similarity in intergroup behaviour". *European Jounral of Social Psychology* 3 (1): 27–52.

Blanco, Mariana, Dirk Engelmann, Alexander K. Koch, and Hans Theo Normann. 2014. "Preferences and beliefs in a sequential social dilemma: a within-subjects analysis". *Games and Economic Behavior* 87:122–135.

Bocian, Konrad, and Bogdan Wojciszke. 2014. "Self-Interest Bias in Moral Judgments of Others' Actions". *Personality and Social Psychology Bulletin* 40 (7): 898–909.

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch. 2014. "hroot: Hamburg Registration and Organization Online Tool". *European Economic Review* 71:117–120.

Bolton, By Gary E, and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition". *American Economic Review* 90 (1): 166–193.

Brown, Rupert J., and John C. Turner. 1979. "The Criss-cross Categorization Effect in intergroup discrimination". *British Journal of Social and Clinical Psychology* 18 (4): 371–383.

Cappelen, Alexander W., Astri Drange Hole, Erik Ø. Sørensen, and Bertil Tungodden. 2007. "The Pluralism of Fairness Ideals: An Experimental Approach". *American Economic Review* 97 (3): 818–827.

Cerrone, Claudia, and Christoph Engel. 2019. "Deciding on behalf of others does not mitigate selfishness: An Experiment". *Economics Letters* 183:108616.

Chen, Daniel L., Martin Schonger, and Chris Wickens. 2016. "oTree—An open-source platform for laboratory, online, and field experiments". *Journal of Behavioral and Experimental Finance* 9:88–97.

Chen, Yan, and Sherry Xin Li. 2009. "Group identity and social preferences". *American Economic Review* 99 (1): 431–457.

Coate, Stephen, and Michael Conlin. 2004. "A Group Rule–Utilitarian Approach to Voter Turnout: Theory and Evidence". *American Economic Review* 94 (5): 1476–1504.

Costa-Gomes, Miguel A., Yuan Ju, and Jiawen Li. 2019. "Role-Reversal Consistency: An Experimental Study of the Golden Rule". *Economic Inquiry* 57 (1): 685–704.

Crippa, M., G. Oreggioni, D. Guizzardi, M. Muntean, E. Schaaf, E. Lo Vullo, E. Solazzo, F. Monforti-Ferrario, J. G. J. Olivier, and E. Vignati. 2019. *Fossil CO2 and GHG emissions of all world countries - 2019 Report*. Tech. rep. Luxemburg: Publications Office of the European Union.

Crisp, Richard J., and Miles Hewstone. 1999. "Differential Evaluation of Crossed Category Groups: Patterns, Processes, and Reducing Intergroup Bias". *Group Processes & Intergroup Relations* 2 (4): 307–333.

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

Davis, Mark H. 1980. "A Multidimensional Approach to Individual Differences in Empathy". *JSAS Catalog of Selected Documents in Psychology* 10:85.

— . 1983. "Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach". *Journal of Personality and Social Psychology* 44 (1): 113–126.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and bounding experimenter demand". *American Economic Review* 108 (11): 3266–3302.

Di Tella, Rafael, Ricardo Perez-Truglia, Andres Babino, and Mariano Sigman. 2015. "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism". *American Economic Review* 105 (11): 3416–3442.

Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York, NY: Harper / Row.

Engelmann, Jan, Maël Lebreton, Peter Schwardmann, Joel J. van der Weele, and Li-Ang Chang. 2019. *Anticipatory Anxiety and Wishful Thinking*. Mimeo.

Enke, Benjamin. 2020. "Moral Values and Voting". *Journal of Political Economy*: forthcoming.

Epley, Nicholas, and Eugene M. Caruso. 2004. "Egocentric Ethics". *Social Justice Research* 17 (2): 171–187.

European Social Survey. 2014. *ESS Round 7 Source Questionnaire*. ESS ERIC Headquarters, Centre for Comparative Social Surveys, City University London, London, United Kingdom.

Exley, Christine L. 2016. "Excusing Selfishness in Charitable Giving: The Role of Risk". *Review of Economic Studies* 83 (2): 587–628.

Exley, Christine L., and Judd B. Kessler. 2019. *Motivated Errors*. NBER Working Paper. Cambridge, MA: National Bureau of Economic Research.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences". *Quarterly Journal of Economics* 133 (4): 1645–1692.

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde. 2016. *The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences*. IZA Discussion Paper 9674. Bonn: Institute for the Study of Labor.

Falk, Armin, and Urs Fischbacher. 2006. "A theory of reciprocity". *Games and Economic Behavior* 54 (2): 293–315.

Feddersen, Timothy, and Alvaro Sandroni. 2006a. "A Theory of Participation in Elections". *American Economic Review* 96 (4): 1271–1282.

— . 2006b. "The calculus of ethical voting". *International Journal of Game Theory* 35 (1): 1–25.

Fehr, Ernst, and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity". *Journal of Economic Perspectives* 14 (3): 159–181.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation". *Quarterly Journal of Economics* 114 (3): 817–868.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Fischbacher, Urs, and Franziska Föllmi-Heusi. 2013. "Lies in Disguise—An Experimental Study on Cheating". *Journal of the European Economic Association* 11 (3): 525–547.

Fischbacher, Urs, Simeon Schudy, and Sabrina Teyssier. 2014. "Heterogeneous reactions to heterogeneity in returns from public goods". *Social Choice and Welfare* 43 (1): 195–217.

Fliessbach, Klaus, Bernd Weber, Peter Trautner, Thomas Dohmen, Uwe Sunde, Christian E. Elger, and Armin Falk. 2007. "Social Comparison Affects Reward-Related Brain Activity in the Human Ventral Striatum". *Science* 318 (5854): 1305–1308.

Gächter, Simon, Daniele Nosenzo, Elke Renner, and Martin Sefton. 2012. "Who Makes a Good Leader? Cooperativeness, Optimism and Leading-by-Example". *Economic Inquiry* 50 (4): 953–967.

Gächter, Simon, Daniele Nosenzo, and Martin Sefton. 2013. "Peer Effects in Pro-Social Behavior: Social Norms or Social Preferences?" *Journal of the European Economic Association* 11 (3): 548–573.

Gerlitz, Jean-Yves, and Jürgen Schupp. 2005. *Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. Dokumentation der Instrumententwicklung BFI-S auf Basis des SOEP-Pretests 2005*. Research Notes 4. Berlin: Deutsches Institut für Wirtschaftsforschung (DIW).

Gino, Francesca, Shahar Ayal, and Dan Ariely. 2013. "Self-serving altruism? The lure of unethical actions that benefit others". *Journal of Economic Behavior and Organization* 93:285–292.

Gino, Francesca, Michael I. Norton, and Roberto A. Weber. 2016. "Motivated Bayesians: Feeling Moral While Acting Egoistically". *Journal of Economic Perspectives* 30 (3): 189–212.

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. 2020. "Bribing the Self'. *Games and Economic Behavior* 120 (311-324).

Graham, Jesse, Jonathan Haidt, and Brian A. Nosek. 2009. "Liberals and Conservatives Rely on Different Sets of Moral Foundations". *Journal of Personality and Social Psychology* 96 (5): 1029–1046.

Güth, Werner, and Hartmut Kliemt. 1998. "The Indirect Evolutionary Approach". *Rationality and Society* 10 (3): 377–399.

Haidt, Jonathan. 2001. *The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment.*

Haidt, Jonathan, and Jesse Graham. 2007. "When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize". *Social Justice Research* 20 (1): 98–116.

Haidt, Jonathan, and Craig Joseph. 2004. "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues". *Daedalus* 133 (4): 55–66.

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

Harsanyi, John C. 1977. "Rule Utilitarianism and Decision Theory". *Erkenntnis* 11 (1): 25–53.

Hippel, Svenja, and Sven Hoeppner. 2019. "Biased judgements of fairness in bargaining: A replication in the laboratory". *International Review of Law and Economics* 58:63–74.

Hofmeier, Jana, and Thomas Neuber. 2019. *Motivated by Others' Preferences? An Experiment on Imperfect Empathy.* CRC TR 224 Discussion Paper 96. Bonn and Mannheim, Germany: Collaborative Research Center Transregio 224 (CRC TR 224).

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

Kant, Immanuel. 1996. "Groundwork of the metaphysics of morals". In *The Cambridge edition of the works of Immanuel Kant: Practical philosophy*, ed. by Mary J. Gregor, 37–108. Cambridge, United Kingdom: Cambridge University Press.

Kassas, Bachir, and Marco A. Palma. 2019. "Self-serving biases in social norm compliance". *Journal of Economic Behavior and Organization.*

Kessler, Judd B., and Stephen Leider. 2012. "Norms and Contracting". *Management Science* 58 (1): 62–77.

Kimbrough, Erik O., and Alexander Vostroknutov. 2016. "Norms Make Preferences Social". *Journal of the European Economic Association* 14 (3): 608–638.

Konow, James. 2001. "Fair and square: the four sides of distributive justice". *Journal of Economic Behavior and Organization* 46 (2): 137–164.

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

— . 2003. "Which Is the Fairest One of All? A Positive Analysis of Justice Theories". *Journal of Economic Literature* 41 (4): 1188–1239.

Kube, Sebastian, Sebastian Schaube, Hannah Schildberg-Hörisch, and Elina Khachatryan. 2015. "Institution formation and cooperation with heterogeneous agents". *European Economic Review* 78:248–268.

Kunda, Ziva. 1987. "Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories". *Journal of Personality and Social Psychology* 53 (4): 636–647.

— . 1990. "The Case for Motivated Reasoning". *Psychological Bulletin* 108 (3): 480–498.

Leeuwen, Boris van, Ingela Alger, and Jörgen W. Weibull. 2019. *Estimating Social Preferences and Kantian Morality in Strategic Interactions.* Mimeo.

Loewenstein, George, Samuel Issacharoff, Colin Camerer, and Linda Babcock. 1993. "Self-Serving Assessments of Fairness and Pretrial Bargaining". *Journal of Legal Studies* 22 (1): 135–159.

López-Pérez, Raúl. 2008. "Aversion to norm-breaking: A model". *Games and Economic Behavior* 64 (1): 237–267.

Malo, Sebastien. 2018. "Marshall Islands marches toward zero greenhouse emissions by 2050". *Reuters*.

Messick, David M., and Keith P. Sentis. 1979. "Fairness and preference". *Journal of Experimental Social Psychology* 15 (4): 418–434.

Mijovic-Prelec, Danica, and Drazen Prelec. 2010. "Self-deception as self-signalling: A model and experimental evidence". *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1538): 227–240.

Mullen, Brian, Rupert Brown, and Colleen Smith. 1992. "Ingroup bias as a function of salience, relevance, and status: An integration". *European Journal of Social Psychology* 22 (2): 103–122.

Oster, Emily, Ira Shoulson, and E. Ray Dorsey. 2013. "Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease". *American Economic Review* 103 (2): 804–830.

Ostrom, Elinor. 2000. "Collective Action and the Evolution of Social Norms". *Journal of Economic Perspectives* 14 (3): 137–158.

Quattrone, George A., and Amos Tversky. 1984. "Causal Versus Diagnostic Contingencies: On Self-Deception and on the Voter's Illusion." *Journal of Personality and Social Psychology* 46 (2): 237–248.

Richerson, Peter J., and Robert Boyd. 2005. *Not By Genes Alone*. Chicago, IL: The University of Chicago Press.

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

Roemer, John E. 2010. "Kantian Equilibrium". *Scandinavian Journal of Economics* 112 (1): 1–24.

Ross, Lee, David Greene, and Pamela House. 1977. "The "False Consensus Effect": An Egocentric Bias in Social Perception and Attribution Processes". *Journal of Experimental Social Psychology* 13 (3): 279–301.

Schwardmann, Peter, Egon Tripodi, and Joël J. van der Weele. 2019. *Self-Persuasion: Evidence from Field Experiments at Two International Debating Competitions*. CESifo Working Paper 7946. Munich, Germany: CESifo.

Schwardmann, Peter, and Joël van der Weele. 2019. "Deception and self-deception". *Nature Human Behavior* 3:1055–1061.

Shafir, Eldar, and Amos Tversky. 1992. "Thinking through uncertainty: Nonconsequential reasoning and choice". *Cognitive Psychology* 24 (4): 449–474.

Slovic, Paul, Melissa Finucane, Ellen Peters, and Donald G. MacGregor. 2002. "Rational actors or rational fools: Implications of the effects heuristic for behavioral economics". *Journal of Socio-Economics* 31 (4): 329–342.

Smith, Megan K., Robert Trivers, and William von Hippel. 2017. "Self-deception facilitates interpersonal persuasion". *Journal of Economic Psychology* 63:93–101.

Sugden, Robert. 1989. "Sugden, Robert". *Journal of Economic Perspectives* 3 (4): 85–97.

Tajfel, Henri, M. G. Billig, and R. P. Bundy. 1971. "Social categorization and intergroup behaviour". *European Journal of Social Psychology* 1 (2): 149–178.

Thompson, Leigh, and George Loewenstein. 1992. "Egocentric Interpretations of Fairness and Interpersonal Conflict". *Organizational Behavior and Human Decision Processes* 51 (2): 176–197.

Turner, J. C., R. J. Brown, and H. Tajfel. 1979. "Social comparison and group interest in ingroup favouritism". *European Journal of Social Psychology* 9 (2): 187–204.

Vanbeselaere, Norbert. 2000. "The Treatment of Relevant and Irrelevant Outgroups in Minimal Group Situations With Crossed Categorizations". *The Journal of Social Psychology* 140 (4): 515–526.

Wilson, Timothy D., and Nancy Brekke. 1994. "Mental Contamination and Mental Correction: Unwanted Influences on Judgments and Evaluations". *Psychological Bulletin* 116 (2): 117–142.

Zajonc, R. B. 1980. "Feeling and Thinking: Preferences Need No Inferences". *American Psychologist* 35 (2): 151–175.

Zellner, Arnold. 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias". *Journal of the American Statistical Association* 57 (298): 348–368.

Zhan, Wei, Catherine C Eckel, and Philip J Grossman. 2020. *Does How We Measure Altruism Matter? Playing Both Roles in Dictator Games*. Mimeo.

Zimmermann, Florian. 2020. "The Dynamics of Motivated Beliefs". *American Economic Review* 110 (2): 337–363.

Appendix A Theoretical Details

A.1 Proofs

Proof of Lemma 1. Choose any $\tilde{c}_{EF}, \tilde{c}_{EQ} \in (1, 20)$ and fix α at a positive value such that the two remaining true fairness parameters that follow from the first-order conditions are also strictly positive.

$$\beta_1 = \frac{\alpha\, Pay'(\tilde{c}_{EF}, role_{EF}) - Inequal'_{EF}(\tilde{c}_{EF})}{Ineff'(\tilde{c}_{EF})},$$
$$\beta_2 = \frac{\alpha\, Pay'(\tilde{c}_{EQ}, role_{EQ}) - Inequal'_{EQ}(\tilde{c}_{EQ})}{Unfair'(\tilde{c}_{EQ})}.$$

Recall that the agent's prior beliefs about the values of the unknown parameters are independently normally distributed with standard deviations of one. The expected values are the true values for β_1 and β_2, while it is $(1-e)\,\alpha$ for α, with $e \in [0, 1]$. Thus, the likelihood of any set of values under the prior beliefs is

$$\mathcal{L} = \phi(\tilde{\alpha} - (1-e)\,\alpha) \times \phi\Big(\tilde{\beta}_1 - \beta_1\Big) \times \phi\Big(\tilde{\beta}_2 - \beta_2\Big),$$

where ϕ denotes the probability density function of the standard normal distribution. The agent maximizes the corresponding log likelihood subject to the first order conditions.

$$\begin{aligned}
&\max_{\tilde{\alpha},\tilde{\beta}_1,\tilde{\beta}_2} \mathit{Constant} - \frac{(\tilde{\alpha} - (1-e)\,\alpha)^2 + \left(\tilde{\beta}_1 - \beta_1\right)^2 + \left(\tilde{\beta}_2 - \beta_2\right)^2}{2} \\
&\text{s.t.} \quad \tilde{\alpha}\, \mathit{Pay}'(\tilde{c}_{EF}, \mathit{role}_{EF}) - \tilde{\beta}_1\, \mathit{Ineff}'(\tilde{c}_{EF}) - \mathit{Inequal}'_{EF}(\tilde{c}_{EF}) = 0 \\
&\qquad \tilde{\alpha}\, \mathit{Pay}'(\tilde{c}_{EQ}, \mathit{role}_{EQ}) - \tilde{\beta}_2\, \mathit{Unfair}'(\tilde{c}_{EQ}) - \mathit{Inequal}'_{EQ}(\tilde{c}_{EQ}) = 0
\end{aligned}$$

In the below notation, derivatives of functions are indicated by small letters and the affective choice as an argument of the functions is omitted. Moreover, define

$$D = \mathit{ineff}^2 \left(\mathit{unfair}^2 + \mathit{pay}(\mathit{role}_{EQ})^2\right) + \mathit{unfair}^2\, \mathit{pay}(\mathit{role}_{EF})^2 \ .$$

Observe that D is always strictly positive. The unique solution of the maximization problem has the following properties.

$$\tilde{\alpha} - \alpha = -\frac{e\,\alpha\, \mathit{ineff}^2\, \mathit{unfair}^2}{D} \tag{A.1}$$

$$\tilde{\beta}_1 - \beta_1 = -\frac{e\,\alpha\, \mathit{ineff}\, \mathit{unfair}^2\, \mathit{pay}(\mathit{role}_{EF})}{D} \tag{A.2}$$

$$\tilde{\beta}_2 - \beta_2 = -\frac{e\,\alpha\, \mathit{ineff}^2\, \mathit{unfair}\, \mathit{pay}(\mathit{role}_{EQ})}{D} \tag{A.3}$$

Part 1 follows from Equation A.1. Parts 2a and 2b follow from Equations A.2 and A.3. □

Proof of Lemma 2. Choose any $\tilde{c}_{EF}, \tilde{c}_{EQ} \in (1, 20)$ and fix α at a positive and γ at a strictly positive value such that the two remaining true fairness parameters that follow from the first-order conditions are also strictly positive.

$$\begin{aligned}
\beta_1 &= \frac{\alpha\, \mathit{Pay}'(\tilde{c}_{EF}, \mathit{role}_{EF}) - \gamma\, \mathit{Inequal}'_{EF}(\tilde{c}_{EF})}{\mathit{Ineff}'(\tilde{c}_{EF})}\,, \\
\beta_2 &= \frac{\alpha\, \mathit{Pay}'(\tilde{c}_{EQ}, \mathit{role}_{EQ}) - \gamma\, \mathit{Inequal}'_{EQ}(\tilde{c}_{EQ})}{\mathit{Unfair}'(\tilde{c}_{EQ})}\,.
\end{aligned}$$

Recall that the agent's prior beliefs about the values of the unknown parameters are independently normally distributed with standard deviations of one. The expected values are the true values for β_1, β_2, and γ, while it is $(1-e)\,\alpha$ for α, with $e \in [0, 1]$. Thus, the likelihood of any set of values under the prior beliefs is

$$\mathcal{L} = \phi(\tilde{\alpha} - (1-e)\,\alpha) \times \phi\left(\tilde{\beta}_1 - \beta_1\right) \times \phi\left(\tilde{\beta}_2 - \beta_2\right) \times \phi(\tilde{\gamma} - \gamma)\ ,$$

where ϕ denotes the probability density function of the standard normal distribution. The

agent maximizes the corresponding log likelihood subject to the first order conditions.

$$\max_{\tilde{\alpha},\tilde{\beta}_1,\tilde{\beta}_2,\tilde{\gamma}} \; Constant - \frac{(\tilde{\alpha}-(1-e)\,\alpha)^2 + \left(\tilde{\beta}_1-\beta_1\right)^2 + \left(\tilde{\beta}_2-\beta_2\right)^2 + (\tilde{\gamma}-\gamma)^2}{2}$$

$$\text{s.t.} \quad \tilde{\alpha}\, Pay'(\tilde{c}_{EF}, role_{EF}) - \tilde{\beta}_1\, Ineff'(\tilde{c}_{EF}) - \tilde{\gamma}\, Inequal'_{EF}(\tilde{c}_{EF}) = 0$$

$$\tilde{\alpha}\, Pay'(\tilde{c}_{EQ}, role_{EQ}) - \tilde{\beta}_2\, Unfair'(\tilde{c}_{EQ}) - \tilde{\gamma}\, Inequal'_{EQ}(\tilde{c}_{EQ}) = 0$$

In the below notation, derivatives of functions are indicated by small letters and the affective choice as an argument of the functions is omitted. Moreover, define

$$\begin{aligned} D = {} & ineff^2 \left(unfair^2 + inequal_{EQ}^2 + pay(role_{EQ})^2\right) \\ & + unfair^2 \left(inequal_{EF}^2 + pay(role_{EF})^2\right) \\ & + \left(inequal_{EF}\, pay(role_{EQ}) - inequal_{EQ}\, pay(role_{EF})\right)^2 . \end{aligned}$$

Observe that D is always strictly positive. The unique solution of the maximization problem has the following properties.

$$\tilde{\alpha} - \alpha = -\frac{e\,\alpha\left[ineff^2\left(unfair^2 + inequal_{EQ}^2\right) + unfair^2\, inequal_{EF}^2\right]}{D} \tag{A.4}$$

$$\tilde{\beta}_1 - \beta_1 = -\frac{\begin{array}{r} e\,\alpha\, ineff\left[\left(unfair^2 + inequal_{EQ}^2\right) pay(role_{EF})\right. \\ \left. - inequal_{EF}\, inequal_{EQ}\, pay(role_{EQ})\right] \end{array}}{D} \tag{A.5}$$

$$\tilde{\beta}_2 - \beta_2 = -\frac{\begin{array}{r} e\,\alpha\, unfair\left[\left(ineff^2 + inequal_{EF}^2\right) pay(role_{EQ})\right. \\ \left. - inequal_{EF}\, inequal_{EQ}\, pay(role_{EF})\right] \end{array}}{D} \tag{A.6}$$

$$\tilde{\gamma} - \gamma = -\frac{e\,\alpha\left(ineff^2\, inequal_{EQ}\, pay(role_{EQ}) + unfair^2\, inequal_{EF}\, pay(role_{EF})\right)}{D} \tag{A.7}$$

Part 1 of the Lemma directly follows from Equation A.4. The results for $\tilde{\gamma}$ of Parts 2a and 2b directly follow from Equation A.7. To see both statements' results for β_1 and β_2, observe that $inequal_{EQ}^2\, pay(role_{EF}) < inequal_{EF}\, inequal_{EQ}\, pay(role_{EQ})$ implies that $inequal_{EF}^2\, pay(role_{EQ}) > inequal_{EF}\, inequal_{EQ}\, pay(role_{EF})$. Thus, for roles (A, a), it cannot hold that $\tilde{\beta}_1 < \beta_1$ and at the same time $\tilde{\beta}_2 < \beta_2$. Conversely, for roles (B, b), it cannot hold that $\tilde{\beta}_1 > \beta_1$ and at the same time $\tilde{\beta}_2 > \beta_2$. Parts 2c and 2d directly follow from Equations A.5 and A.6. □

A.2 Hypothesis Testing

Denote by $A_{i,g}$ an indicator variable taking the values zero and one for whether subject i from group g would personally profit if subjects from group $g-1$ chose higher rather than lower options $C^1_{\cdot,g-1}$ according to Procedure 1. Similarly, denote by a an indicator for whether she would profit from higher options $C^2_{\cdot,g-1}$ in the case of Procedure 2. We

estimate the following two regressions:

$$C^1_{i,g} = \delta_0 + \delta_1 A_{i,g} + \epsilon_{i,g}$$
$$C^2_{i,g} = \zeta_0 + \zeta_1 a_{i,g} + \eta_{i,g}$$

We want to conduct the following statistical hypothesis test:

$$H_0: \quad \delta_1 \leq 0 \vee \zeta_1 \leq 0$$
$$H_1: \quad \delta_1 > 0 \wedge \zeta_1 > 0$$

Thus, we want to reject the Null hypothesis of either coefficient being weakly negative, i.e., we want to establish that both coefficients are strictly positive.

Note that $A_{i,g}$ and $a_{i,g}$ are statistically independent, since all combinations or roles appear with exactly the same frequencies in the experiment. Moreover, $\epsilon_{i,g}$ and $\eta_{i,g}$ are each pairwise statistically independent of both $A_{i,g}$ and $a_{i,g}$, since assignment to roles is randomized.

To understand the implications of the above discussion for the hypothesis test, consider the following scenario: we have estimated the two regression equations from above and retrieved p-values p_δ and p_ζ referring to the two-sided significance tests of δ_1 and ζ_1, respectively. The p-value referring to the above hypothesis test is the probability of either of the two t-values under H_0 (t^0_δ and t^0_ζ) being as large as they are (t_δ and t_ζ), with at least one of δ_1 and ζ_1 being smaller than zero.

$$\begin{aligned}
p &= P\big(t^0_\delta \geq t_\delta \wedge t^0_\zeta \geq t_\zeta \,\big|\, \delta_1 \leq 0 \vee \zeta_1 \leq 0\big) \\
&= P(\delta_1 \leq 0 \,|\, \delta_1 \leq 0 \vee \zeta_1 \leq 0) \times P\big(t^0_\delta \geq t_\delta \wedge t^0_\zeta \geq t_\zeta \,\big|\, \delta_1 \leq 0\big) \\
&\quad + P(\zeta_1 \leq 0 \,|\, \delta_1 \leq 0 \vee \zeta_1 \leq 0) \times P\big(t^0_\delta \geq t_\delta \wedge t^0_\zeta \geq t_\zeta \,\big|\, \zeta_1 \leq 0\big) \\
&\quad - P(\delta_1 \wedge \zeta_1 \leq 0 \,|\, \delta_1 \leq 0 \vee \zeta_1 \leq 0) \times P\big(t^0_\delta \geq t_\delta \wedge t^0_\zeta \geq t_\zeta \,\big|\, \delta_1 \leq 0 \wedge \zeta_1 \leq 0\big) \\
&\leq P\big(t^0_\delta \geq t_\delta \wedge t^0_\zeta \geq t_\zeta \,\big|\, \delta_1 \leq 0\big) + P\big(t^0_\delta \geq t_\delta \wedge t^0_\zeta \geq t_\zeta \,\big|\, \zeta_1 \leq 0\big) \\
&\leq P\big(t^0_\delta \geq t_\delta \wedge t^0_\zeta \geq t_\zeta \,\big|\, \delta_1 = 0 \wedge \zeta_1 \to \infty\big) + P\big(t^0_\delta \geq t_\delta \wedge t^0_\zeta \geq t_\zeta \,\big|\, \delta_1 \to \infty \wedge \zeta_1 = 0\big) \\
&= P\big(t^0_\delta \geq t_\delta \,\big|\, \delta_1 = 0\big) + P\big(t^0_\zeta \geq t_\zeta \,\big|\, \zeta_1 = 0\big) \\
&= \frac{p_\delta + p_\zeta}{2}
\end{aligned}$$

The p-value referring to the hypothesis test is thus the average of the two standard p-values from the OLS regressions.

Appendix B Empirical Details

Table B.1: Sample Composition

	Obs.	Mean	Median	Min.	Max.
Age	372	25.583	24	18	72
Female	369	0.599	1	0	1
University student	372	0.836	1	0	1
player.income	372	741.185	600	0	3500
ln_income	368	6.283	6.39693	0	8.160519

Table B.2: Nominal Group Bias in Decisions

Dependent variable	*Decision*		*Belief*	
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Same name is A	2.086***		0.958	
	(0.731)		(0.590)	
Same name is a		2.441***		1.715***
		(0.687)		(0.513)
Constant	7.454***	4.457***	6.935***	4.118***
	(0.503)	(0.437)	(0.417)	(0.300)
Observations	372	372	372	372
R^2	0.022	0.033	0.007	0.029

Note: Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B.3: Nominal Group Bias in Decisions (with Roles)

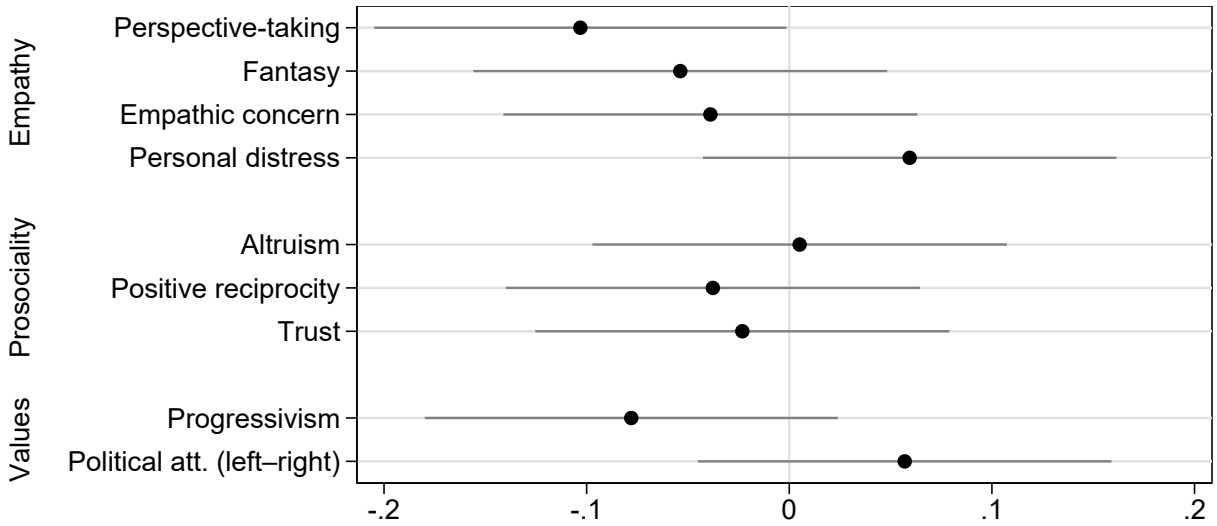
Dependent variable	*Decision*		*Belief*	
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Same name is A	2.086***		0.958	
	(0.720)		(0.583)	
Same name is a		2.441***		1.715***
		(0.669)		(0.502)
Role A	2.602***		1.849***	
	(0.720)		(0.583)	
Role a		3.140***		2.102***
		(0.669)		(0.502)
Constant	6.160***	2.887***	6.016***	3.067***
	(0.617)	(0.439)	(0.500)	(0.367)
Observations	372	372	372	372
R^2	0.055	0.087	0.033	0.073

Note: Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B.4: Order Effects

Dependent variable	*Decision*		*Belief*	
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Role A	3.576***		2.502***	
	(1.006)		(0.825)	
Role $A \times EQ$ first	-1.946		-1.292	
	(1.453)		(1.166)	
Role a		2.603***		2.808***
		(1.001)		(0.712)
Role $a \times EQ$ first		1.107		-1.427
		(1.357)		(1.020)
EQ first	1.170	-1.048	0.885	0.727
	(0.992)	(0.848)	(0.778)	(0.598)
Constant	6.596***	4.615***	6.034***	3.573***
	(0.688)	(0.662)	(0.544)	(0.410)
Observations	372	372	372	372
R^2	0.038	0.058	0.030	0.049

Note: Heteroscedasticity-consistent standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.



Note: Figure shows the Pearson correlation coefficient for the ENA proxy introduced in Equation 5 and the respective survey measure. Gray bars indicate 95% confidence intervals.

Figure B.1: Correlations with *ENA* Proxy (Full Sample)

Table B.5: Heterogeneity (Full Sample)

Dependent variable	*ENA proxy*			
	(1)	(2)	(3)	(4)
Perspective-taking	-0.152**		-0.151**	-0.158**
	(0.0700)		(0.0706)	(0.0709)
Fantasy	-0.0320		-0.0378	-0.0337
	(0.0655)		(0.0665)	(0.0681)
Empathic concern	0.0102		0.0276	0.0278
	(0.0703)		(0.0769)	(0.0776)
Personal distress	0.161**		0.166**	0.170**
	(0.0681)		(0.0685)	(0.0694)
Altruism		-0.0148	0.00246	0.00503
		(0.0578)	(0.0642)	(0.0646)
Positive reciprocity		-0.0320	-0.0183	-0.0141
		(0.0522)	(0.0527)	(0.0538)
Trust		-0.0463	-0.0570	-0.0541
		(0.0568)	(0.0556)	(0.0561)
Personality controls	Yes	Yes	Yes	Yes
Demographic controls	No	No	No	Yes
Observations	372	372	372	372
R^2	0.075	0.044	0.078	0.083

Note: The table reports standardized coefficients. Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. The *personality controls* are risk and time preferences along with the Big Five traits. The *demographic controls* are gender (categories: female, male, and diverse), age and squared age, a dummy for being enrolled at university, and the log of monthly gross income in euros, calculated as $\ln(income + 1)$.