ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 116

Egocentric Norm Adoption

Thomas Neuber

September 2021

www.econtribute.de

UNIVERSITÄT BONN

Universität zu Köln

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – EXC 2126/1– 390838866 is gratefully acknowledged.

Egocentric Norm Adoption*

Thomas Neuber†

September 10, 2021

Abstract

Social norms pervade human interaction, but their demands are often in conflict. To understand behavior, it is thus crucial to know how individuals resolve normative tradeoffs. This paper proposes that sincere judgments about the relative importance of conflicting norms are shaped by personal interest. We show that people tend to follow norms from which they benefit themselves, even in contexts where their own decisions only affect others. In a (virtual) laboratory experiment, each subject makes two decisions over allocations of points within a group of two other participants. The sets of possible allocations entail different normative tradeoffs, and subjects have no personal stakes in their own decisions. However, they are affected by others' decisions: each subject is part of a group, and the members of different groups simultaneously decide over others' allocations along a circle. We find that subjects' decisions are biased towards the normative principles aligned with their own interests, thereby favoring other players whenever these share those interests. Subjects' beliefs about the choices made by others suggest a largely unconscious mechanism. Moreover, survey answers indicate that the effects are driven by self-centered reasoning: subjects who report pronounced perspective-taking are less biased.

Keywords: egocentrism, experiment, social norms
JEL codes: C91, D63, D91

---

*This paper has benefited from discussions with Thomas Dohmen, Raphael Epperson, Armin Falk, Jana Hofmeier, Philipp Strack, Axel Wogrolly, Florian Zimmermann, and Christian Zimpelmann. Helpful comments have also been received from seminar participants at CERGE-EI, the University of Exeter, the Norwegian School of Economics (NHH), and Universitat Pompeu Fabra (UPF), as well as at the ESEM Winter Meeting 2020 and the EEA Annual Congress 2021. 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. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 – 390838866.

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

1 Introduction

People care about adhering to social norms, but different norms are often in conflict.[1] Due to opposing prescriptions, it is unclear in many situations what constitutes appropriate and fair behavior. The economic literature has considered this issue from two different angles. One has been to elicit people's *true* attitudes regarding specific tradeoffs (Konow, 2003; Cappelen et al., 2007), often using *impartial spectators* who decide as third parties over allocations between others (Konow, 2000, 2009; Cappelen et al., 2013). The other approach has been to study how people decide about normative tradeoffs when they are affected by their own choices. It has been found that people exploit "moral wiggle room" to excuse selfish behavior (Dana, Weber, and Kuang, 2007). Thus, the two existing approaches either mute self-interest or introduce it directly. However, in many economically relevant situations, an indirect channel might be important: personal interest shapes normative views and is thereby even relevant when, in a particular situation, there are no incentives to behave selfishly.

This paper proposes that people tend to follow norms aligned with personal interest, even when their own actions do not secure them any advantage. Consider a court case and an unprejudiced judge who neither personally knows any involved party nor has any personal interest in the matter under review. However, the judge shares a certain case-relevant feature with one of the parties, e.g., being male in the context of gender discrimination. Since women's interests are—presumably—less relevant for the judge's personal welfare than those of men, his decision might be biased. Similarly, corporate leaders perhaps think what their staff policies would have meant for themselves at earlier stages of their careers and—perhaps unconsciously—are therefore reluctant towards affirmative action policies. In both cases, people make decisions affecting others that reflect what kind of general behavior is beneficial for themselves, apparently because personal interest has shaped their relative support for different norms. For this phenomenon, we introduce the term *egocentric norm adoption*.

In applied settings, people's interests are correlated with various characteristics, and the potential repercussions of actions are often complex. To provide clean evidence for egocentric norm adoption, we designed a laboratory experiment with three central features: First, subjects are affected by others' choices over normative tradeoffs. Second, subjects' interests are exogenously varied, i.e., they are randomly allocated to roles that profit or lose from certain norms. Third, they also decide in the same decision contexts themselves but over others, such that they are not affected by their own decisions. Specifically, pairs of subjects are randomly assigned to groups. For the two members of each group, subjects from other groups choose allocations of points. The possible allocations involve tradeoffs between two different fairness norms, where each of the principles favors

[1] For the general importance of social norms in economics, see, e.g., Elster (1989) and Ostrom (2000).

one of the group members. Subjects simultaneously 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. The experiment consists of two decision contexts: the *EF Procedure* trades-off equality against efficiency,[2] while the *EQ Procedure* involves equality and equity, i.e., the principle that divisions of a surplus should reflect individual contributions. Subjects have distinct roles for each procedure that determine from which respective normative principles they profit, and the roles of subjects in adjoining groups are crossed. Before making any decisions, each subject knows that she 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.[3] Thus people tend to follow norms from which they would personally benefit if they were adhered to by *others*. Alger and Weibull (2013) have argued that from an evolutionary perspective, such behavior should be expected. They have also drawn a connection to Kant's categorical imperative. However, the behavior of subjects in our experiment seems to follow intuition rather than principled reasoning. After 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 self-centered reasoning driving the results, we find that decisions are less biased among subjects who report pronounced perspective-taking.

Throughout their lives, people gain or lose depending on the prevalence of various normative principles. Hence, egocentric norm adoption suggests that people living under different circumstances develop different normative views. Therefore, it can potentially explain some of the heterogeneity in decisions made by impartial spectators, or what Cappelen et al. (2007) call the "pluralism of fairness ideals." Consider, e.g., the subjects that Cappelen et al. classify as *libertarian*, who believe that even *random* productivity differences should be reflected in payoffs. Perhaps, these individuals have adopted this normative view because they have benefited themselves from random events outside the experimental context. This reasoning is supported by the finding that, among a sam-

[2]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).

[3]The term *bias* here refers to systematic differences in subjects' behavior with no normative justification. A different approach would be to define bias relative to some normative benchmark. That could be the average decision of impartial spectators (see Konow, 2000, 2009; Cappelen et al., 2013) or subjects that are part of the same experiment but uninformed about their own roles.

ple of adolescents in Norway, high-socioeconomic status (SES) spectators exhibited less egalitarianism than their low-SES counterparts (Almås et al., 2017).

How the concept of egocentric norm adoption can potentially explain economically relevant attitudes can be seen in greater detail from three stylized facts about support for public redistribution. (i) Support for national redistribution is decreasing in family income, as Alesina and Giuliano (2011) show with data from the World Value Survey (WVS). This relationship is found even though most people have virtually no individual power over political decisions, implying that they have no economic motives for self-deception. (ii) Using US data from the General Social Survey, the same article also finds a negative association between support for national redistribution and family income when the respondents were 16 years old, conditional on current family income. The fact that attitudes persist when interests change indicates that they are genuine. Attitudes towards redistribution appear to be influenced by personal interest, but induced shifts can even show in (temporal) contexts where they are unconnected to self-interest. (iii) Support for foreign aid among people in donor countries is *increasing* in income, as Chong and Gradstein (2008) show with data from the WVS. Thus, while the rich and the poor favor their likes concerning national redistribution, the picture is reversed for global redistribution. The above pattern can neither be satisfyingly explained by plain self-interest nor by group cohesion due to socioeconomic status. However, egocentric norm adoption delivers a parsimonious explanation for all three findings: people hold genuine normative views that are more than excuses for selfishness, but their views are nonetheless guided by personal interest. People who are poor within their countries support more national redistribution because they would benefit themselves. They are truly convinced of their normative views and stick with them even if their own situation changes. However, the poor in a rich country support less *global* redistribution, as they suspect an outflow of resources that would otherwise be spent on them.

The experiment's results suggest a certain behavioral mechanism that underlies the phenomenon of egocentric norm adoption: people are ill-equipped to fully abstract from what decisions like their own would imply for themselves. Hence, their capability to empathize with interests different from their own is limited. This mechanism explains why the effects are also present in beliefs and why they decrease in perspective-taking, i.e., people's tendency to "put themselves in others' shoes." The psychological literature has noted that people who are in a given emotional state find it difficult to predict reactions of themselves or others in different emotional states (see Van Boven et al., 2013). The implications of such egocentric empathy gaps have been explored by Van Boven, Dunning, and Loewenstein (2000) in the context of the endowment effect. People who own an object get "attached" to it, and they project their heightened valuations upon potential buyers. Regarding wider economic questions, however, it appears that the topic

has received virtually no attention.[4] This paper is part of a research agenda to explore the economic implications of egocentric reasoning. A related paper by Hofmeier and Neuber (2019) is concerned with how people's willingness to help depends on how much they would appreciate the same kind of 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. Subjects are thus imperfectly empathic in acting not only upon receivers' preferences but also upon their own.

The experiment presented in the current paper stresses the negative side of egocentric norm adoption, i.e., its egocentric aspect. As discussed above, the mechanism likely adds to explaining disagreement about fairness standards and distributive policies, even between people who personally are unaffected and could thus claim to be impartial. However, there is also (or, perhaps, primarily) a positive message: people seem to engage in normative reasoning and adopt subjectively demanded behavior. In the related 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. Similarly, egocentric norm adoption might have positive consequences in many social situations and, in particular, promote cooperation between individuals with shared interests. For example, it could 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 could help overcome collective action problems and supply public goods because people in such situations share the same interests.[5] This insight also has practical implications for effective communication in the face of collective action problems. During the current COVID-19 pandemic, e.g., an important policy goal is convincing people to wear face masks, which deliver more protection to people around the wearer than to the wearer herself. In light of this paper's findings, it would be promising to stress people's self-interest in *others* wearing face masks. Realizing their own stakes, people should consider the norm of wearing masks important and more readily comply with it themselves.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 then introduces in detail the experimental design. The derivation of the hypotheses follows in Section 4. Section 5 presents the main results. Subsequently, Section 6 conducts an analysis of heterogeneity in the observed effects. Finally, Section 7

[4]For a general discussion of why emotions should be given a more prominent role in the economic literature, see Elster (1998). For the particular relevance of empathy, see also Singer and Fehr (2005).

[5]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).

summarizes the paper discusses the results.

2 Related Literature

The present 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 by oneself and 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 evidence for cognitive dissonance reduction (Festinger, 1957; Akerlof and Dickens, 1982).[6] 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), uncertainty about the prevalence of a given norm (Bicchieri, Dimant, and Sonderegger, 2020), sharing the benefits of unethical 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, biases in fairness views are induced by direct monetary incentives. 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

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

the field setting of a debating competition.[7]

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 own 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.[8] This paper agrees and shows that egocentric perceptions of potential outcomes do not just affect how people feel about narrowly-defined 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 interest" (see the examples in Section 1).

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

[7]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 only 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.

[8]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.

groups overlap, there is an interest in effects from 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 given 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. In another economic lab experiment, Cassar and Klein (2019) show that group identity can also be induced by common experiences of success or failure, leading to corresponding favoritism in decisions over redistribution.

Our paper relates to this literature in that egocentric norm adoption can give rise to a phenomenon akin to in-group–out-group bias. People treat others well if they share the same economic interests. If economic interests in a particular situation coincide among some groups of people and differ for others, discrimination arises between "interest groups." The experiment rules out classical in-group–out-group bias by crossing roles between adjoining groups. Subjects know that both group members for whom they choose an allocation are in one of their in-groups and one of their out-groups, such that SIT would not make any prediction for differential treatment.[9] Moreover, the crossing of roles implies that egocentric norm adoption favors a different participant for each of the two decisions that a subject makes.

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 opts for strategies that would benefit her if everyone else also adopted them. Roemer (2010, 2015) shows that in the presence of externalities, equilibria arising from Kantian maximization dominate Nash equilibria. He concludes by arguing that evolution might have made Kantian behavior common in humans due to group selection (Roemer, 2015). Indeed, using a different formal approach and assuming assortative matching of interacting individuals, Alger and Weibull (2013) show that evolution should converge to a mixture of selfish and Kantian preferences.[10] 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 struc-

[9]However, players in the experiment also have *names* X and Y in each group, which are independent of roles (see Section 3). SIT predicts bias in favor of players sharing subjects' own names, which we find in Section 5.3.

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

turally 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 high probability as the second mover.[11] As has also 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 driven by principled philosophical reasoning, as the reference to Immanuel Kant would suggest, but mainly unconscious, which is confirmed by our finding of biased beliefs. Empirically, we do not rely on interpreting individual-level patterns in behavior but are the first to use the aspect of egocentrism. Identification relies on exogenously induced interests—i.e., on roles—, and egocentric norm adoption is thereby 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 as it turns out, a realistic characterization of people's intuitive behavior.

3 Experiment

People constantly lose or benefit from different normative principles, and egocentric norm adoption predicts that this shapes their normative views. In the experiment, we randomly vary which principles align with subjects' personal interests or are opposed to them. These manipulations are small regarding subjects' overall lives, but they are salient during the experiment. Thus, they allow for a causal test of whether personal interest influences adherence to different norms.

People first learn about their own group and their personal interest in the two allocation procedures, i.e., their roles. It is made salient from the beginning of the experiment that they cannot influence their own payoffs. Next, they are informed about the details of the group for which they decide. After everything has been firmly understood, subjects make their two decisions. These are followed by the elicitation of beliefs about other subjects' choices, and the experiment concludes with several questionnaires. The

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

full translated instructions are available from the author's website.[12]

3.1 Design

A multiple of four participants takes part in each experimental session. Pairs of participants are randomly allocated to groups, numbered consecutively from 1 to N. In each group, one participant is called *Player X*, and the other participant's name is *Player Y*. All participants receive a fixed participation fee of €4 and, during the experiment, points each worth €0.01. Importantly, no player makes any decision regarding their own group. Instead, groups simultaneously decide over players in other groups along a circle, i.e., Group 1 decides over Group 2, Group 2 decides over Group 3, ..., and Group N decides over Group 1. Every player makes two decisions over allocations of points for the players in the respective succeeding group, each according to a different procedure. One decision is about the tradeoff between equality and efficiency (EF Procedure); the other is about the tradeoff between equality and equity, i.e., attribution of responsibility (EQ Procedure). For the EF Procedure, one player in each group takes the role that profits from efficiency, while the other player profits from equality. In the paper, we denote the former role by A and the latter by B. For the EQ Procedure, we denote roles by a and b, where Role a profits from equity and Role b from equality. The labels of roles do not appear in the instructions, and they are determined independently of subjects' names (X and Y). The instructions do not use the labels for the procedures, either. Instead, these are called "Procedure 1" and "Procedure 2," depending on their randomly determined order on the subject level. Any two players in any two adjoining groups share exactly one role. Figure 1 visualizes this structure, where tuples after players' names denote their roles in the EF and the EQ Procedure, respectively.

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

Figure 1: Example for Roles in Successive Groups

Estimation Task The EQ Procedure requires that subjects can contribute to the success of their groups. Therefore, all subjects have to engage in an estimation task. The task precedes all other instructions of the experiment, and we tell subjects that a precise estimate will increase their chances of receiving additional money during the experiment. On their computer screens, subjects see a three-second countdown, after which they see an image for two seconds. The image shows a certain number of blue dots on a yellow background. Immediately after the image has disappeared, subjects have 15 seconds to

[12] https://thomas-neuber.github.io/papers/ENA_instructions.pdf

enter an estimate for the number of dots that they saw. Their task is to minimize the absolute difference between their estimate and the actual number of dots.[13] Before the actual task, subjects complete an identical trial task with a different number of dots. The respective images that subjects see are the same for all participants, showing 40 dots for the trial task and 53 for the actual task. Neither of these numbers is revealed to subjects.

After the estimation task, subjects learn about the experiment's basic setup, i.e., the circular decision structure. The instructions spell out precisely who makes decisions concerning the group to which they belong themselves and for which group they will make decisions. A highlighted box emphasizes that they will in no way be able to influence the allocation of points within their own group. Players first learn about names and roles within their own group and the potential payoff consequences for themselves and their partners. Afterward, they are informed about the structure of the group for which they decide. This order mimics typical real-life situations in which people know about their interests (e.g., being rich or poor) before considering a particular decision problem (voting over a redistributive policy measure).

Efficiency (EF) Procedure The EF procedure concerns the tradeoff between equality in points for both individual players and efficiency regarding the total number of points. The possible allocations of points are shown in Table 1.

Table 1: Payoffs for the EF 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 in the choice 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. Then, it 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. Thus, relative to the fully equal outcome, efficiency can be increased by a factor of up to 2.5. However, efficiency gains decrease from 90 points between Option 1 to Option 2 to just a

[13]The task of estimating the number of dots follows the one used in Fliessbach et al. (2007). However, the original task 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, thereby avoiding ties.

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 beginning of the experiment, all players engaged in an estimation task, which they were told would increase their chances of getting additional money (see above). The estimates that subjects gave are used for the EQ Procedure in which 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 estimate of the player in Role b does not affect how many points the group receives.[14] 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 EQ 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 that reflect accountability for the total points that the group received, i.e., a reward for the player who won the points.

The instructions display the potential payoffs like Tables 1 and 2, except that participants see the names of players (X and Y) instead of roles. The row for Player X is always on top, and that for Player Y below, i.e., the two rows might be reversed.[15] Subjects have to correctly answer three sets of control questions while reading through the instructions, for which they can reread the relevant previous screens. The first set of questions follows the information about their own group. These questions refer to the experiment's structure and roles in the subjects' own groups. For two example options, subjects have to fill in the amounts of points that both players in their group would receive. A corresponding second set of questions is presented after subjects have learned about the situation within the groups over which they decide. The last set of control questions regards the crossed roles between groups and the below information about the implementation of payoffs.

[14]The fact that subjects cannot learn about their performance and that everybody took part in the same task under the same conditions mutes any self-esteem motives.

[15]The fixed and transparent order facilitates understanding. Subjects find their own payoff in the same row for both procedures. By favoring their own roles, players once give advantage to the subject sharing their own row and once to the subject whose payoff is displayed in the other row.

Afterward, subjects make their decisions for the respective succeeding group, one after the other in the subjective-specific order. No option is preselected.

At the end of the experiment, the computer conducts a three-step random procedure to implement a subset of decisions. First, it randomly chooses one of the two procedures. Second, it determines whether decisions come from either all even- or all odd-numbered groups. Third, it determines one subject within each relevant group and implements their respective decisions. Thus, for 50% of subjects, a decision made by another subject is implemented. The 25% of subjects whose own decisions become relevant themselves receive 1,000 points.[16] For the remaining 25% of subjects, their payoff depends on another task independent of their own decisions (see the paragraph on belief elicitation below).

Belief Elicitation After the two decisions, we elicit players' beliefs about choices by others. Specifically, we ask them to guess the average of the choices that subjects from other groups in their session have made for groups that, in terms of the role compositions, are identical to the one for which they have decided themselves. If the decision of a subject's group partner is implemented, i.e., with a probability of 25%, the guess's accuracy determines their payoff. Average choices within the same session are calculated for each procedure, separately for even- and odd-numbered groups, and excluding each subject's own group.[17] Subjects then receive 500 points if their guess is precisely correct and 250 points as long as the correct answer falls into the range of the five options closest to their guess. We elicit the beliefs with tables that look exactly like the ones for the decisions. The tables highlight the range of options for which the currently selected option would still imply 250 points.[18]

Questionnaires The experiment proceeds with a survey asking subjects about fundamental sociodemographic characteristics like age, gender, and income. Moreover, participants complete several questionnaires on personality, preferences, and values. The details with the corresponding results are presented in Section 6. Finally, subjects learn their payoffs and the details of how they came about.

[16]For these subjects, 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 design properties alleviate concerns that subjects' decisions over others might depend on expectations about their own payoffs, e.g., due to aversion towards inequality (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*, we should observe a negative correlation between choices and beliefs. As Section 5.2 will show, the opposite is the case.

[17]This procedure makes sure that the relevant other subjects decided over groups that, abstracting from players' names (who was X and who was Y), are identical to the one for which the respective participant was deciding herself. It also ensures that the roles of comparison subjects are balanced, i.e., that the different roles are present in equal numbers.

[18]For options towards either end of the scale, the interval for which subjects receive 250 points becomes asymmetric around the reported belief. This asymmetry ensures that subjects whose true beliefs are at the extremes have no mechanic incentive to adjust their answers towards the center.

Let us conclude this description of the experimental design by pointing out two noteworthy features that allow for a clean identification of egocentric norm adoption. First, the experiment's structure 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). Second, the experimental design comprises two different procedures, such that each player has two roles. Own roles and roles of subjects for whom players decide are crossed, and players thus know that they share exactly one role with each subject over which they decide. Thereby, we distinguish the effect of egocentric norm adoption from in-group–out-group bias in the sense of SIT. According to the latter, preferential treatment is due to elevated attitudes towards in-group members relative to out-group members. Such reasoning is focused on *others*, and it would take into account both of a subject's two roles (hence the interest in how people aggregate crossed categorizations; see Section 1). If both procedures were equally relevant for identity, SIT would predict no effect. If one procedure were more important than the other, SIT would predict that a given player favors the same subject in both decisions. In contrast, when people egocentrically adopt norms, they are not focused on others but *themselves*. Preferential treatment is not attached to other people but an individual's roles. Therefore, in the experiment, egocentric norm adoption predicts that a given player favors a different subject for each procedure, i.e., always the one who shares the player's respective own role.

3.2 Implementation

The experiment was run from May 13 until May 20, 2020, and implemented as a virtual lab experiment. Seventeen sessions with either 20 or 24 subjects resulted in a total of 372 participants who completed the experiment.[19] Participants were recruited from the subject pool of the *BonnEconLab* using the software *hroot* (Bock, Baetge, and Nicklisch, 2014). The experiment's language was German, and we invited only German-speaking subjects. Participants were mostly university students, and around 60% of subjects were women. For details of the sample composition, see Table B.1 in the appendix. 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.[20] They received individual links, such that it was impossible for any subject to participate more than once. Since we ran the experiment during the first phase of the COVID-19 pandemic, subjects presumably participated from home (the university library, e.g., was closed at the time). Contrary to typical online experiments, however,

[19]We had to exclude four of the 376 participants who initially started the experiment because they either stopped working on the experiment or were unable to answer some of the control questions.

[20]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.

and just as in a usual laboratory experiment, subjects attended specific experimental sessions. They had to participate in the experiment at a pre-specified time and date. Other participants in the same session were taking part simultaneously, and an experimenter was available to answer questions. On the introduction screen, we gave subjects contact details which they could use in case of questions. The experimenter was available via email, telephone, or text.[21] Subjects had already received the contact details before the experiment as part of the automated email communications (invitation, an email with the personal link, reminder). Several subjects asked questions during the experiment, and all contact methods were used.

4 Hypotheses

The paper's main hypothesis is that participants make decisions favoring their own role for the respective procedure. To understand the reasoning behind this conjectured effect in the absence of material incentives or, in fact, *any* instrumental or otherwise self-serving motives, we develop a simple formal framework that attributes biased fairness views not to motivated cognition but the (partial) inability to abstract from one's own role. The framework is inspired by Haidt (2001), who argues that people commonly make ethical judgments based on intuitive reactions and that moral reasoning often takes the form of mere ex-post rationalization. Building on this insight, Epley and Caruso (2004) have conjectured that intuitive moral evaluation in conjunction with automatic egocentrism can explain self-serving ethical judgments. The framework presented here offers a way of formalizing the existing arguments and makes a conceptual contribution by shifting the focus from specific *judgments* to beliefs about generally applicable *norms*. This novel perspective is critical for the resulting behavioral implications: only if people attribute their self-centered affective reactions to the relative importance of norms, the egocentric bias carries over to decisions that do not personally affect them.

Consider, e.g., a metaphor from soccer. A player from a team that a given person supports commits a foul. The intuitive reaction of the supporter is that "this was not a foul." She will perhaps come up with reasons for her judgment, which could take various forms. She could question inferences drawn from video evidence or accuse the opposing player of diving. This kind of reasoning would not affect her judgment of situations between any other teams. However, she could also come to believe that the rules should be changed and that more physical play should be generally permitted. This *would* change her judgment of other situations. In this paper, we suggest that personal interest changes how people think about the "rules of the game," and not just about situational factors. The formal framework will assume that they rationalize their affective reactions *exclusively*

[21] In contrast to using an online conference platform, these contact methods allowed for one-to-one communication between subjects and the experimenter.

by changing the perceived importance of generally applicable norms, which is the limiting case that makes the argument most transparent. Based on our theoretical conjecture, we derive the testable behavioral implication of egocentric norm adoption.

4.1 Formal Framework

The starting point of the formal framework is that, while considering the possible choice options, an agent experiences an affective reaction determined by her fairness views but also by the payoff implied for her own relevant role because her perspective is inherently subjective: options implying a high payoff for herself "feel good." Her fairness views and level of subjectivity are, however, imperfectly known to the agent. Instead, she knows which option yields the most positive affective reaction. When being confronted with the choice that she has to make over others, she tries to empathize with those affected. She thus engages in the underlying normative tradeoff and tries to learn about the importance of the involved norms.[22] 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. However, if she is affected by some degree of egocentrism, i.e., her ability of perspective-taking is imperfect, she underestimates the influence of subjectivity. She arrives at fairness-views that depend on her own roles and at corresponding choices that are egocentrically biased.

4.1.1 Basic Setup

The agent makes one choice for the EF and one for the EQ Procedure, c_{EF} and c_{EQ}, respectively. We assume that the choice set for both procedures is the interval $[1, 20]$, i.e., the agent can choose intermediate options. When considering a given option for one of the procedures, the agent experiences an affective reaction depending on her own respective payoff and the violation of the two norms that are relevant in the respective procedure. We denote by $role_{EF} \in \{A, B\}$ the agent's role in the EF Procedure and by $role_{EQ} \in \{a, b\}$ her role in the EQ Procedure. The agent's affective reaction functions for the two procedures are given by the following equations.

$$React_{EF}(c_{EF}) = \alpha\, Pay(c_{EF}, role_{EF}) - \beta_1\, Ineff(c_{EF}) - Inequal_{EF}(c_{EF}) \tag{1}$$

$$React_{EQ}(c_{EQ}) = \alpha\, Pay(c_{EQ}, role_{EQ}) - \beta_2\, Unfair(c_{EQ}) - Inequal_{EQ}(c_{EQ}) \tag{2}$$

The 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

[22] If the agent did not want to exert any effort at all to make her decisions, she would choose randomly.

in the choice while, for Options B and b, it is strictly decreasing. Thus, it may simply correspond to the number 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. 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 first order conditions for Equations 1 and 2. The set of solutions is not atomic, and different combinations of parameters can rationalize the intuitive optimum. For example, a high value of $\tilde{c}_{EF}$ 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 the beliefs about β_1 and β_2, the means are the respective true values, while for belief about α, the mean is multiplied by $\pi \in [0, 1]$. The latter parameter denotes the level of perspective-taking. It captures the ability to recognize how affective reactions depend on roles. The prior belief is given by $\mathcal{N}(\pi\alpha, 1)$.[23] Her decision-relevant beliefs are the values that are most likely given her prior beliefs and the two first-oder conditions.

Lemma 1. *Assume positive subjectivity ($\alpha > 0$) and limited perspective-taking ($\pi < 1$). Then:*

1. *The agent underestimates her level of subjectivity, i.e., $\tilde{\alpha} < \alpha$.*
2. *The agent's updated fairness views 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. □

Lemma 1 formally captures the intuition of egocentric norm adoption: an agent who profits from efficiency-oriented decisions by others will tend to consider this normative principle important. In contrast, an agent who personally loses from efficient allocations will tend to object to the principle. Similarly, an agent who profits from equity-oriented decisions will support the corresponding principle more strongly than an agent who loses from them.

[23] 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. The results on perspective-taking in Section 6 support the interpretation in the sense of perspective-taking.

4.1.2 Combining the Procedures

The above framework considers the decisions for the two procedures independently, which suffices for the main predictions. Note, however, that both the EF Procedure and the EQ 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 and b always profits from equality, while one with roles Roles B and a or Roles A and b profits from equality according to one procedure and loses in the other. Lastly, the private interest of a participant with roles Roles A and a is always opposed to equality. Using this feature, the setup allows for insights into how egocentrically adopted norms can spill over from their source to other contexts. Formally, let us modify Equations 1 and 2 in the following way:

$$React_{EF}(c_{EF}) = \alpha\, Pay(c_{EF}, role_{EF}) - \beta_1\, Ineff(c_{EF}) - \gamma\, Inequal_{EF}(c_{EF}) \tag{3}$$

$$React_{EQ}(c_{EQ}) = \alpha\, Pay(c_{EQ}, role_{EQ}) - \beta_2\, Unfair(c_{EQ}) - \gamma\, Inequal_{EQ}(c_{EQ}) \tag{4}$$

In contrast to the previous assumptions from Equations 1 and 2, the agent now also forms a belief about the importance of equality, γ. This creates a connection between the two procedures regarding the relative importance of the involved norms, just as it has been intuitively discussed above. As before, the agent knows her true reaction functions up to the now four parameters. All beliefs are the same as before, and the prior belief about γ also follows a Normal distribution with a standard deviation of one, centered around the true value. From the modified assumptions, the below results follow.

Lemma 2. *Assume positive subjectivity ($\alpha > 0$) and limited perspective-taking ($\pi < 1$). Then:*

1. *The agent underestimates her level of subjectivity, i.e., $\tilde{\alpha} < \alpha$.*
2. *The agent's updated 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. □

Lemma 2 states that an agent who always loses if others are taking equality-oriented decisions applies a weight to equality that is biased downward. Moreover, there is an upward bias in at least one of the weights that she assigns to the opposing norms, i.e., efficiency and equity. The opposite is true for an agent who always gains from equality.

We can make no statement about the weight attached to equality for agents who gain from equality-oriented decisions in one procedure and lose from them in the other. However, for the respective opposing norm involved, the same applies as in Lemma 1: agents who benefit from decisions that emphasize efficiency or equity consider the respective norms important, while they otherwise exhibit a downward bias in the attached weight.

4.2 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 4.1.1, the objective functions that she *wants* to maximize are Equations 1 and 2, setting the value of α to zero. 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 4.1.1, the objective functions are concave. Under these conditions, the agent's choices c^*_{EF} and c^*_{EQ} are uniquely identified by the following first-order conditions.

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

Both optima's locations 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 roles.*

To test the hypothesis formally, denote by $r_i^{EF} \in \{A, B\}$ the role of subject i for the EF Procedure and by $r_i^{EQ} \in \{a, b\}$ her role for the EQ Procedure. The subject's choice for the EF Procedure is denoted by $c^*_{i,EF}$ and the one for EQ Procedure by $c^*_{i,EQ}$. Hypothesis 1 was preregistered, and in the pre-analysis plan we committed to running the two following regressions:

$$c^*_{i,EF} = \delta_0 + \delta_1 1_A\big(r_i^{EF}\big) + \epsilon_{i,g} \tag{5}$$
$$c^*_{i,EQ} = \zeta_0 + \zeta_1 1_a\big(r_i^{EQ}\big) + \eta_{i,g} \tag{6}$$

The terms $1_A\big(r_i^{EF}\big)$ and $1_a\big(r_i^{EQ}\big)$ denote indicator functions for roles A and a, respectively. Since subjects in roles A and a would profit from higher-ordered choices by the respective sending group, egocentric norm adoption predicts that both δ_1 and ζ_1 should be positive. Note that the hypothesis requires *both* coefficients to be positive. In Appendix A.2, we show that an upper bound for the joint one-sided p-value is provided by the average of the

separate two-sided p-values. This result means that if both coefficients are significantly positive in separate OLS regressions, the null hypothesis of either coefficient being weakly negative can be rejected.

An agent who loses from equality in both procedures (i.e., whose roles are A and a) will initially feel attracted to high choice options. As has been shown in Lemma 2, she will view this as strong evidence that she cares little about equality and a lot about 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 in one procedure and lose from it in the other one will notice that their initially preferred choices are somewhat contradictory as one seems to reflect strong concern about equality while the other does not. These observations lead to the following hypothesis.

Hypothesis 2. *Among participants 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 expect spillovers of roles to the respective other decision contexts, i.e., a positive effect of Role A on the decision for the EQ Procedure and, similarly, a positive effect of Role a on the choice for the EF Procedure.

In the formal framework introduced here, the bias in choices arises unconsciously and 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. We thus have a further hypothesis.

Hypothesis 3. *Similarly to decisions, beliefs about others' decisions are biased in favor of subjects' respective roles.*

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

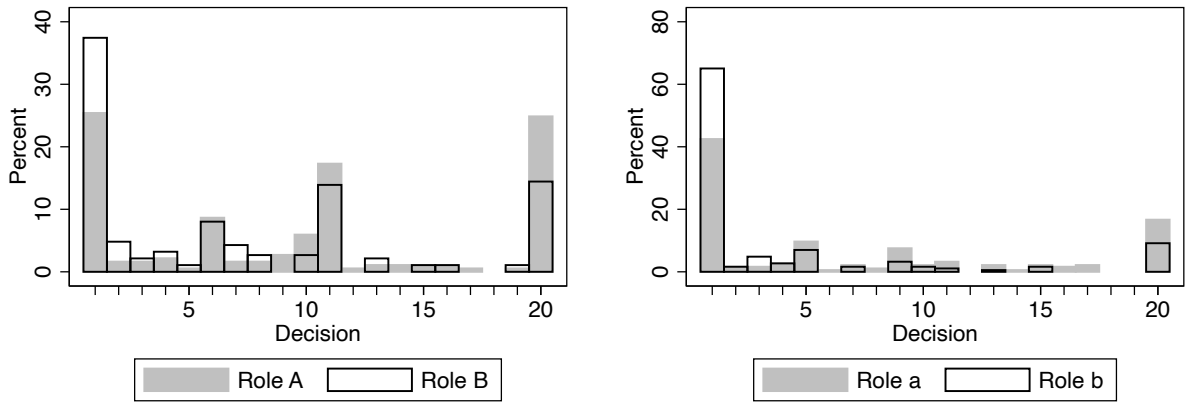
5 Main Results

This section presents the main results of the experiment, starting with the decisions in Section 5.1 and proceeding with the analysis of beliefs in Section 5.2. Section 5.3 discusses further observations.

5.1 Decisions

Figure 2 visualizes the decisions that subjects made in the experiment. The two panels are identically constructed. The left one displays the distribution of decisions for the

EF Procedure, and the right panel shows the decisions for the EQ Procedure. In displaying the distributions of decisions, the panels differentiate between the two relevant roles for the respective procedure. For the EF procedure, these are Role A (shaded), profiting from efficiency, and Role B (light), profiting from equality. 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

Notes: 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 others choosing high options choose higher options themselves 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 columns show the estimates for the central regressions equations, i.e., Equations 5 and 6. 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 identical and, in quantitative terms, slightly

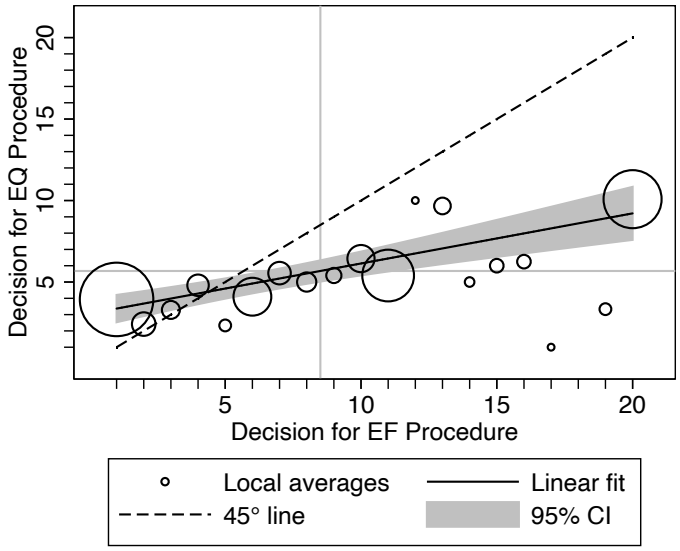
stronger. Here, the median option chosen by subjects in Role a is 5, while it is 1 for subjects in Role b. The means are 7.25 and 4.11, respectively. The difference between the latter values corresponds to 0.47 standard deviations and is thus even larger than the one observed for the EF Procedure. Column 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*			
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Role A	2.602***		2.609***	1.304*
	(0.727)		(0.725)	(0.677)
Role a		3.140***	1.224*	3.147***
		(0.680)	(0.725)	(0.677)
Constant	7.209***	4.108***	6.593***	3.456***
	(0.498)	(0.428)	(0.585)	(0.475)
Observations	372	372	372	372
R^2	0.033	0.055	0.041	0.064

Notes: 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 both procedures involve the criterion of equality, once weighted against efficiency (EF Procedure), and the other time against the equity principle (EQ Procedure). Suppose a subject puts a strong emphasis on equality. In that case, this should manifest itself in low choices for both procedures. On the other hand, one would expect a subject who does not consider equality to be important to make high choices for both procedures. Thus, choices for the two procedures should be positively correlated among subjects. Figure 3 displays the empirical relationship between the two decisions that subjects are making. For every option for the EF Procedure on the horizontal axis, the vertical axis shows the respective players' average decisions for the EQ Procedure. The sizes of circles correspond to the relative number of subjects. We observe a clear positive trend. The upward-sloping regression line confirms the positive relationship. It is based on the disaggregated data and corresponds to a correlation of 0.33 ($p < 0.001$, two-sided). The correlation cannot be due to roles since those are independent.



Notes: The figure groups subjects by their decisions for the EF Procedure. For each option on the horizontal axis, the figure plots the respective subjects' average decisions for the EQ Procedure on the vertical axis. The sizes of circles correspond to the respective numbers of subjects. The dashed line indicates 45 degrees. The gray lines indicate the averages of decisions for the EF Procedure (vertical) and the EQ Procedure (horizontal). The solid black line represents the linear fit from an OLS regression, and the shaded area around it corresponds to the 95% confidence interval based on heteroscedasticity-consistent standard errors.

Figure 3: Relationship Between the Two Decisions

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. Columns 3 and 4 of Table 3 again consider the EF Procedure and the EQ Procedure, respectively, but they 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.[24] 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 similarly to preexisting differences between different individuals.[25] Individually, both spillover coefficients are weakly statistically significant ($p < 0.1$), and they are jointly significant at

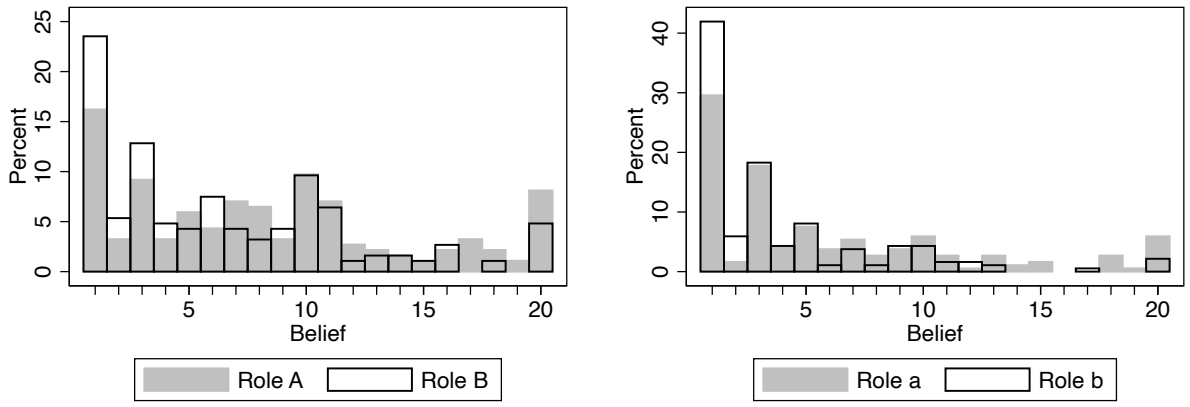
[24]The point estimates slightly differ because, as mentioned earlier in Footnote 19, four subjects did not complete the experiment, and roles are therefore not precisely independent anymore. The slight empirical correlations between roles are random, and the implications for estimates minimal.

[25]Multiplying the effect from Column 1 in Table 3 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.

the five percent level ($p = 0.02$).[26]

5.2 Beliefs

An important question is whether subjects' egocentric behavior is conscious or 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) EF Procedure (b) EQ Procedure

Notes: 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.

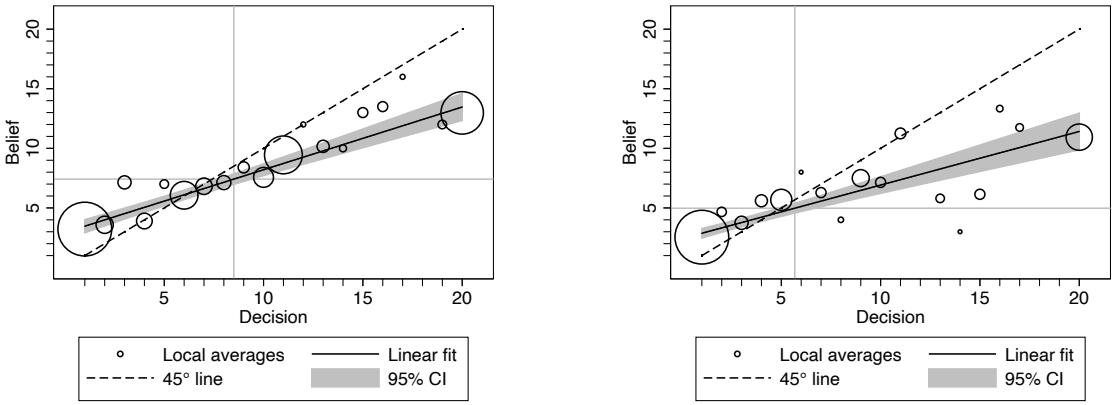
Figure 4: Beliefs by Roles

Figure 4 shows the beliefs about others' choices that subjects reported in the experiment. It is constructed in the same way as Figure 2 for decisions, differentiating between subjects' roles. For both procedures, the distributions of beliefs feature more weight at values around the center than observed for choices. However, concerning role differences, 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. Regression results confirm the visual impression. Role A's effect for the EF Procedure and Role a's for the EQ Procedure are both positive and statistically significant ($p < 0.01$ and $p < 0.001$; see Table B.2 in the appendix). The effects amount to 0.32 standard deviations for the EF Procedure 0.42

[26] In the appendix, we provide a visual decomposition of the spillover effects by distinguishing between the four possible combinations of roles that subjects might have. In the EF Procedure, the spillover effect is mainly driven by subjects who gain from equality in both procedures and choose very low options (see Figure B.2). Subjects who have to hope for equality attach little relative weight to efficiency. In the EQ Procedure, the spillover effect arises symmetrically: it is driven by subjects who profit from equality in both procedures choosing low options and subjects who lose from equality in both procedures choosing high options (see Figure B.3).

standard deviations for the EQ Procedure, which are relative sizes similar to those found for decisions, although slightly smaller.

The results for beliefs resemble those for decisions also in other respects. For decisions, Figure 3 has established a strong positive correlation of 0.33 between the two procedures. The corresponding relationship between subjects' beliefs is also clearly positive ($p < 0.001$, two-sided; see Figure B.1 in the appendix), although the correlation coefficient only amounts to 0.18. In light of this finding, it is not surprising that the spillover effects in beliefs are again positive but smaller and not statistically significant (see Table B.2 in the appendix).



(a) EF Procedure (b) EQ Procedure

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

Figure 5: Decisions and Beliefs

The above results might suggest that the effects of roles on decisions arise unconsciously, at least to a large extent. This interpretation presupposes that beliefs and decisions are positively related—which indeed they are. Figure 5 plots average beliefs for subjects depending on their decisions and is otherwise constructed just like Figure 3. The left panel shows the relationship for the EF Procedure, and the right panel does the same for the EQ Procedure. First, we observe that average beliefs, indicated in both panels by gray lines, are lower than average decisions for both the EF Procedure and the EQ Procedure ($p < 0.001$ and $p = 0.01$, respectively). That means subjects, on average, expect others to assign a higher relative weight to equality than they do themselves. More importantly, we see 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. These associations between decisions and beliefs are instances of the well-established false consensus

effect (Ross, Greene, and House, 1977): people have a fundamental disposition to believe that others' convictions are more similar to their own than they really are. The first two columns of Table 4 confirm that the effect is strong, regressing decisions on beliefs for the EF and the EQ Procedure, respectively. The estimates for both slope parameters are larger than 0.8.

Table 4: Decisions and Beliefs

Dependent variable	*Decision for succeeding group*			
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Belief (EF Procedure)	0.820***		0.805***	
	(0.0511)		(0.0519)	
Belief (EQ Procedure)		0.809***		0.777***
		(0.0695)		(0.0722)
Role *A*			1.115**	
			(0.565)	
Role *a*				1.506**
				(0.606)
Constant	2.417***	1.654***	1.981***	1.058***
	(0.469)	(0.386)	(0.530)	(0.404)
Observations	372	372	372	372
R^2	0.432	0.364	0.438	0.376

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

Thus, we have seen that roles affect beliefs similarly to decisions and that beliefs and decisions are closely related. It suggests an unconscious channel: people who are influenced by their own roles might not realize how they are biased. To the extent that this is the case, the effects of roles on decisions should be reduced once we control for subjects' respective beliefs. The results are shown in Columns 3 and 4 of Table 4. We first note that the coefficients for beliefs are hardly changed. However, the effects of roles change decisively. Recall that, without controlling for beliefs, the effects were 2.602 for the EF Procedure and 3.140 for the EQ Procedure (see Table 3). Now, they are reduced by 57% ($p < 0.01$, two-sided) and 52% ($p < 0.001$), respectively. The remaining effects of roles remain statistically significant ($p = 0.049$ and $p = 0.013$, respectively). However, because our measures of beliefs almost certainly contain some measurement error, the true conditional effects should tend to be even smaller (see Gillen, Snowberg, and Yariv, 2019). In sum, the point estimates suggest that roles' effects on decisions arise to more than one-half unconsciously, and it might even be the case that subjects are entirely unaware of the bias in their decisions.

5.3 Further Observations

The experiment allows for some further noteworthy insights. One is 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 different between affected groups in which one of the players has Roles A and a (*parallel* groups) and other (*crossed*) groups. Within crossed groups, high choices offset each other from an ex-ante perspective because both subjects profit from inequality in one procedure and lose from it in the other. For parallel groups, however, ex-ante inequality from high choices cumulates, since one subject profits from inequality in *both* procedures, while the other loses in both. Thus, one could expect that decisions over players in crossed groups, i.e., those made by subjects in parallel groups, should generally be higher and more strongly positively correlated. Empirically, however, there is no indication of any such differences. The distributions of decisions do not significantly differ between the group types ($p = 0.28$ for the EF Procedure and $p = 1.00$ for the EQ procedure; Kolmogorov–Smirnov test, two-sided). Moreover, the positive correlation between decisions exists for both group types separately ($p < 0.001$ and $p < 0.01$, respectively), and there is no evidence for a difference in the correlations' magnitudes ($p = 0.32$). Despite these similarities in the overall distributions of decisions between the two group types, the effects of players' own roles on decisions are stronger in parallel groups than in crossed groups. In fact, precisely these differences identify the effects of roles across procedures in Columns 3 and 4 of Tables 3 and B.2. In light of the above discussion, which excludes concerns about ex-ante fairness as an alternative explanation, the respective coefficients seem interpreted best as evidence for spillover effects.

By the experiment's design, roles do not induce differential proximity between players in adjoining groups due to crossed roles. However, independently of roles, 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. In line with 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, see Columns 1 and 2 of Table B.3 in the appendix). The effect sizes are smaller than the ones estimated for roles, although the differences are not significant. For beliefs, the corresponding effects are similar but weaker. Both estimated coefficients are positive, and the EQ procedure's effect is significant ($p < 0.001$, see Columns 3 and 4 of

Table B.3). Since names were determined independently of roles, the estimated effects of roles and names are virtually unaffected by including the respective regressors jointly (see Table B.4).[27]

A potential concern in many experiments involving human subjects is experimenter demand. It denotes the possibility that subjects try to conform to the experimenters' expectations. This experiment's design mitigates such concerns to the largest possible extent. An important design property is that treatment effects are identified between subjects, as opposed to within-subjects. The between-subject design avoids making subjects 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 as large as observed in our data. 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 are also present for beliefs, which we elicited with an incentivized procedure. Here, subjects would have had to give up their own money to conform to expectations. The data also generally seem well-behaved. For example, Table B.5 in 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. Lastly, the next section will show that the treatment effects are not mainly due to a few subjects making extreme decisions but caused by the bulk of subjects exhibiting moderate bias.

6 Heterogeneity

This section aims to relate the observed bias induced by roles to relevant personal attributes of subjects. In terms of understanding the mechanism behind our main results, we are particularly interested in the role of perspective-taking. We also consider the role of different aspects of empathy as well as prosociality. Regarding outcomes, we study the relationship between progressivism and political orientation on the left–right spectrum.

[27] In the appendix, we visually inspect the interaction between players' own roles and nominal groups. In the EF Procedure, prominently high options are chosen by subjects whose own role is A and who are in a nominal group with another subject in Role A (see Figure B.4). In the EQ Procedure, the effects of roles and nominal groups appear to be independent (see Figure B.5). Perhaps it matters for these results that payoffs are linear in choice options for the EQ Procedure but not the EF Procedure.

6.1 Attributing the Effects to Subjects

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. Therefore, we first convert each subject's two decisions into a single individual-specific *ENA* proxy of egocentric norm adoption and construct a corresponding measure for biased beliefs in the same way. This proxy intuitively measures how a given subject contributes to the observed treatment effects for decisions. We start from the self-evident fact that, if there were no treatment effects, 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 with 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.

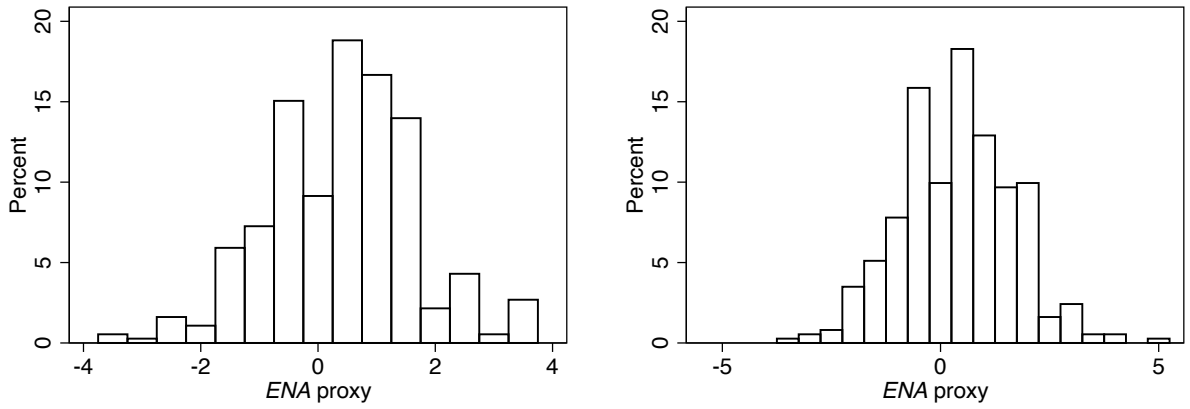
Therefore, we calculate for every decision its deviation from the average of choices for the respective procedure. For better comparability across procedures, we further divide the differences by the respective standard deviation, i.e., we transform subjects' choices into z-scores denoted by z_i^{EF} and z_i^{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 $1_A\left(r_i^{EF}\right)$ for whether subject i's role for the EF Procedure is A and the analogously defined indicator functions for the three remaining procedure–role combinations, we thus calculate the *ENA* proxy as follows:

$$ENA_i \equiv \left[1_A\left(r_i^{EF}\right) - 1_B\left(r_i^{EF}\right)\right] z_i^{EF} + \left[1_a\left(r_i^{EQ}\right) - 1_b\left(r_i^{EQ}\right)\right] z_i^{EQ} \tag{7}$$

Deviations that are aligned with a subject's relevant role contribute to higher values of the *ENA* proxy, 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 making 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 and b, values closer to zero if her roles are A and a or B and b, and a large negative value of the *ENA* proxy if her roles are B and a.

Of course, the decisions of subjects in part also reflect their true fairness convictions. As can be seen from the above examples, however, the expected effect of any given true fairness preferences on the value of the *ENA* proxy is exactly zero. That is because subjects' roles determine the signs that Equation 7 attaches to the z-scores, and roles are drawn randomly with equal probabilities. The *ENA* proxy thus consists of two components: one is any systematic bias in decisions due to roles, and the other is random noise due to subjects' true fairness convictions. The latter is orthogonal to any subject-specific

attributes by construction, while the former might correlate with subjects' personal characteristics.



(a) Decisions (b) Beliefs

Notes: The figure shows the distribution of the *ENA* proxies, one the left for decisions and on the right for beliefs. The respective values have been calculated according to Equation 7 for the full sample.

Figure 6: Distribution of the *ENA* proxies

Figure 6 shows the distributions of the *ENA* proxies for decisions and beliefs, respectively. For decisions, the 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 5.1). This positive average is significantly different from zero ($p < 0.001$, two-sided t-test), in line with the previous findings. 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. Instead, it is also caused by many subjects exhibiting moderate levels of bias towards 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).[28] The picture looks similar for beliefs. Here, the mean value is 0.37 (the average of 0.32 and 0.42; see Section 5.2), which is again significantly different from zero ($p < 0.001$, two-sided t-test). As for decisions, it also holds for beliefs that the average value is still significantly positive among moderate values on the interval $[-1.5, 1.5]$ ($p = 0.018$, two-sided t-test).

6.2 Survey Measures

After the main experiment, subjects completed several questionnaires that were selected to measure potentially relevant personal characteristics. Below, we introduce the elicited classes of characteristics.

[28] In particular, the restriction excludes all subjects who favor their own roles to the largest possible extent, i.e., those with Roles A and a choosing Option 20 for both procedures, with Roles A and b choosing Option 20 for the EF Procedure and Option 1 for the EQ Procedure, with Roles B and a choosing Options 1 and 20, respectively, and with Roles B and b choosing Option 1 for both procedures.

Empathy 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). The IRI measures perspective-taking 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 in need. The above dimensions of empathy are truly directed at others' feelings, and we would expect that they tend to decrease egocentric bias. In contrast, the fourth dimension of *personal distress* is "self-oriented" (Davis, 1983, p. 114) and addresses whether people feel anxious themselves when they witness others' suffering. Batson, Fultz, and Schoenrade (1987) argue that "empathic distress" is a vicarious feeling that is, in fact, distinct from empathy. In terms of behavior, empathy in its altruistic form facilitates helping (Batson et al., 1981), whereas distress induces an egoistic desire for relief. Therefore, personal distress might be expected to increase egocentric bias.

Prosociality The experiment in this paper aims to show a bias that speaks of egocentrism. In contrast, the design mutes the role of egoism with the absence of selfish incentives. However, to study prosociality's role empirically, we included the qualitative item for altruism, positive reciprocity, and trust from the Preference Survey Module (Falk et al., 2016; Falk et al., 2018).

Values 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.[29] As suggested by the developers, we

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

aggregate the five subscales into a single measure of *progressivism*.[30]

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

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 ($r = -0.51, p < 0.001$). Conceptually, we consider the above measures of values as potential *outcomes* of egocentric norm adoption. In contrast, the personality traits of empathy and prosociality are plausible *determinants*.

Personality Controls As control variables, we include the qualitative preference items by Falk et al. (2016) for risk preferences, time preferences, and negative reciprocity. 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 cooperative behaviors; and *neuroticism*, referring to emotional instability and anxiety.

Demographic Controls The elicited sociodemographic controls are subjects' gender (female, male, or diverse) and age, enrollment at a university, and gross monthly income in euros. The latter is converted into log income as $\ln(income + 1)$.

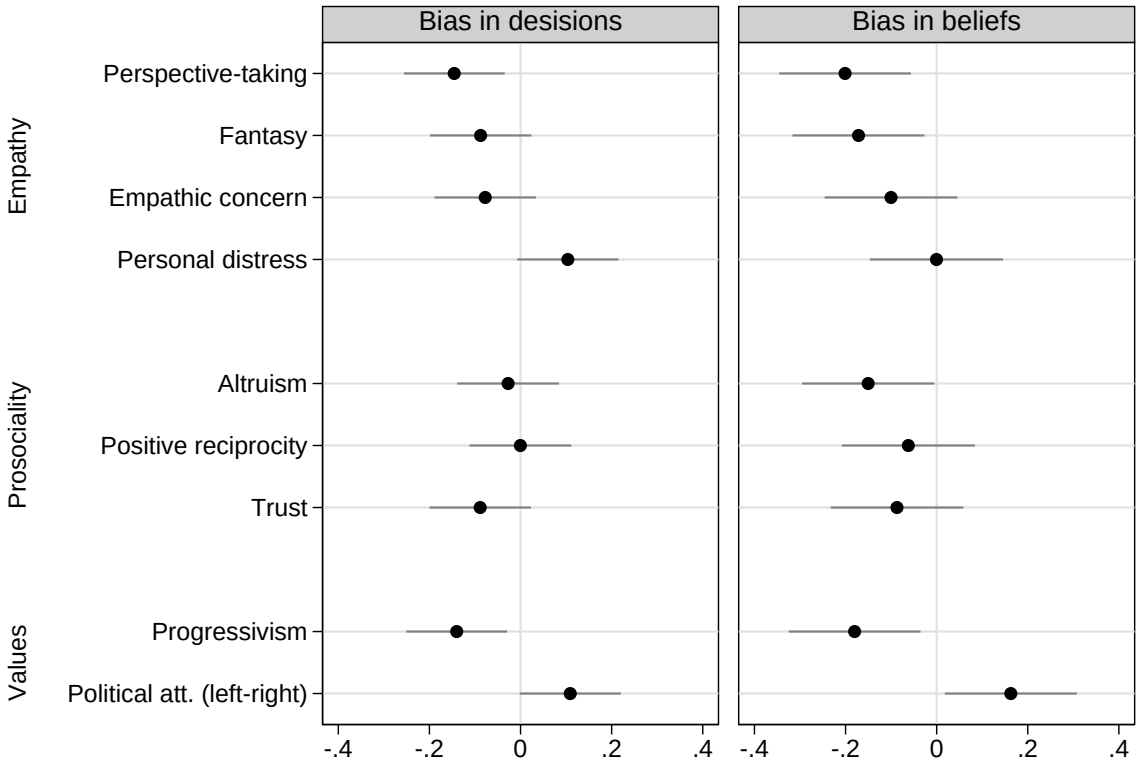
Studying heterogeneity in subjects' behavior is more demanding than the previous analyses in Section 5 for two reasons. First, individual subjects' behavior has stronger effects on results since the sample is, intuitively, split, and the resulting subsamples are smaller than the full one. Second, we use answers to unincentivized survey questions, which some subjects might not have taken very seriously or had difficulty answering. Therefore, we restrict the sample to individuals who had no major difficulties understanding the experiment and took answering the survey questions seriously. The experiment included several control questions, and subjects could only progress once they answered them correctly. 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

[30] 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 our data, the correlation between these two measures based on the same questionnaire is 0.96.

subjects who gave answers below the tenth percentile. These restrictions leave us with 312 subjects, for whom all previous results from Section 5 replicate. Results for the full sample are provided in Appendix B.[31]

6.3 Heterogeneity in Bias

Figure 7 considers the correlations between the *ENA* proxies and the different measures of empathy, prosociality, and values. The left panel shows the correlations between the



Notes: The figure shows the Pearson correlation coefficient between the *ENA* proxies and different survey measures. 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 7: Correlations with the *ENA* Proxies

survey measures and egocentric bias in decisions. For the four dimensions of empathy, a pattern arises that is consistent with the theoretical predictions: the three "altruistic" 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 "egoistic" side of empathy—personal distress—leads to a stronger egocentric bias. The correlation with perspective-taking,

[31] For the construction of the *ENA* proxy, we still use the z-scores based on the full sample. Alternatively, they can be calculated separately for the restricted sample. The corresponding values of the *ENA* proxy are almost identical ($\rho > 0.99$), and all results that follow remain virtually unchanged.

which is the opposite of egocentrism, is significantly negative ($p = 0.01$, two-sided), and the correlation with personal distress is (weakly) significantly positive ($p = 0.07$). The correlations with fantasy and empathic concern are not statistically significant ($p > 0.1$). We do not observe a significant correlation with either of the prosociality measures ($p > 0.1$). 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 weaker egocentric bias than conservatives. Consistent with this finding, people leaning to the political right show a stronger bias than those leaning to the left ($p = 0.05$).[32]

The panel on the right displays the correlations with the *ENA* proxy for beliefs. Overall, they are remarkably similar to decisions. Again, the correlations with the three other-oriented dimensions of empathy are negative. In this case, they are statistically significant for perspective-taking ($p < 0.01$, two-sided) and also for fantasy ($p = 0.02$). The correlation with empathic concern is insignificant ($p > 0.1$). Thus, people who report little perspective-taking are not only more biased than others, but they also project their bias upon others. This finding suggests that perspective-taking, or the lack thereof, occurs unconsciously. It is in line with the assumptions in the formal framework (see Section 4.1) and with the insights provided by Singer and Fehr (2005). Other than observed for decisions, there is no indication of a relationship between bias in beliefs and personal distress. The correlations between the *ENA* proxy for beliefs and the prosociality measures tend to be negative. For one of them—altruism—it is now also statistically significant ($p = 0.04$). The correlations with progressivism and political orientation are very similar to those found for decisions, and they are both statistically significant ($p = 0.01$ and $p = 0.03$, respectively).

In the full sample, correlations of survey measures with the *ENA* proxies for decisions and beliefs are quite similar to those in the restricted sample (see Figure B.6 in the appendix). Importantly, the correlations with perspective-taking remain statistically significant. Other results lose their statistical significance, most likely due to more noise in data. Only the positive correlation between the *ENA* proxy for beliefs and subjects' political attitude remains (weakly) statistically significant.

The analysis of heterogeneity is admittedly 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. Therefore, we employ a regression framework. All of the reported regressions

[32]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).

Table 5: Heterogeneity

Dependent variable	*ENA proxy*			
Domain	*Decisions*		*Beliefs*	
	(1)	(2)	(3)	(4)
Perspective-taking	-0.176**	-0.185**	-0.187**	-0.180*
	(0.0754)	(0.0754)	(0.0923)	(0.0936)
Fantasy	-0.0999	-0.0886	-0.132	-0.136
	(0.0760)	(0.0786)	(0.0934)	(0.0978)
Empathic concern	-0.00779	-0.0109	0.103	0.119
	(0.0783)	(0.0792)	(0.107)	(0.107)
Personal distress	0.248***	0.261***	0.131	0.123
	(0.0712)	(0.0719)	(0.0986)	(0.0991)
Altruism	-0.00313	-0.00408	-0.161*	-0.166*
	(0.0713)	(0.0718)	(0.0876)	(0.0868)
Positive reciprocity	0.0400	0.0473	-0.0349	-0.0196
	(0.0601)	(0.0611)	(0.0800)	(0.0802)
Trust	-0.111*	-0.104*	-0.0961	-0.0877
	(0.0606)	(0.0618)	(0.0761)	(0.0768)
Personality controls	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes
Observations	312	312	312	312
R^2	0.120	0.130	0.093	0.106

Notes: 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 $\ln(income + 1)$.

include the full set of survey measures previously considered in Figure 7. Moreover, all columns control for other personality characteristics (see Section 6.2 above), and we report standardized coefficients. The independent variable in the first column is the *ENA* proxy for decisions. The results confirm the results from the correlations. Perspective-taking is associated with less biased decisions ($p = 0.02$) and personal distress with an increase in bias ($p < 0.01$). Otherwise, only the coefficient for trust is (weakly) statistically significant, entering with a negative sign ($p = 0.07$). The results hardly change when also controlling demographic characteristics in Column 2. The coefficients for all control variables can be found in Table B.6 in the appendix. Columns 3 and 4 replicate the previous two for beliefs. The results for perspective-taking are remarkably similar to those for decisions, emphasizing the interpretation of unconsciousness. As already indicated by the respective correlations, personal distress does not appear to be associated with bias in beliefs. The only other coefficients that are (weakly) statistically significant are the negative ones for altruism.

Overall, this section's main result is that perspective-taking seems to play a central role in the emergence of egocentric norm adoption. Among subjects who report high levels of perspective-taking, the bias in decisions is significantly reduced. Subjects reporting little perspective-taking do not only take more biased decisions themselves, but they also project their own bias upon others. Egocentric norm adoption arises unconsciously, and whether individuals overcome it seems to depend on whether they can abstract from their own perspective.

7 Conclusion

This paper has provided experimental evidence for the phenomenon of egocentric norm adoption. If people would benefit from others following certain norms, they adopt these principles themselves. The experiment's central property 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. Subjects within groups of two players received points according to two procedures. One of these implied a tradeoff between equality and efficiency, and the other involved the norms of equality and equity. Depending on their respective roles, subjects personally gained from one of the norms involved in a procedure and lost from the other. The players of each group decided over the subjects in the respective succeeding groups along a circle. Players' roles (in norms favoring them) were crossed, and players knew that they shared exactly one role with each subject over which they decided. We found an egocentric bias for both procedures and corresponding biases of similar size also in subjects' beliefs about others' behavior. The heterogeneity analysis provides additional support for egocentrism as the critical driving force behind the treatment effects of roles: the bias is largest among subjects who report weak perspective-

taking.

Future research on egocentric norm adoption could explore additional potential mechanisms that might underly the phenomenon. This paper has made the case that an unconscious egocentric bias leads subjects to empathize more with positions that they are in themselves. Our view is supported by the effect arising largely unconsciously and the role of perspective-taking. However, the presence of one mechanism does not rule out the existence of others. The biases in decisions could, in part, also result from subjects confusing *diagnostic* with *causal* contingencies (Quattrone and Tversky, 1984; Shafir and Tversky, 1992; Acevedo and Krueger, 2005; Krueger and Acevedo, 2007), whereby subjects would try to "induce" a desired behavior by others with their own actions. Similarly, the results for decisions and beliefs could, to some extent, reflect "wishful thinking" (see, e.g., Mijovic-Prelec and Prelec, 2010; Engelmann et al., 2019)—although the evidence seems to suggests that this phenomenon is not present for purely financial stakes (Barron, 2020). Future work could adapt this paper's experimental design but put some subjects into a position where a non-human random device like a computer determines their own payoffs. If subjects' decisions partly also reflect a direct concern with others' choice behavior, the egocentric biases should become smaller.

For methodology, the paper's findings and those by Hofmeier and Neuber (2019) caution against the equivalent use of elicitation procedures for social preferences with or without role uncertainty. Iriberri and Rey-Biel (2011) and Zhan, Eckel, and Grossman (2020) find increased prosociality in (modified) dictator games when it is ex-ante uncertain whether a given subject will be paid as dictator according to one of her own decisions or as a receiver according to a decision by another subject. Egocentric norm adoption can accommodate these findings,[33] and it implies more. Whenever subjects play multiple games within an experiment, researchers who want to avoid bias should be aware that interests can induce norms, potentially creating spillovers between different contexts.

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). Thereby "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.* The latter describes the seemingly irrational behavior of people who

[33]Grech and Nax (2020) theoretically and empirically analyze the related but more complex difference between standard, non-interactive dictator games with certain roles and interactive dictators games. In the latter, roles are not uncertain, but subjects have two roles, simultaneously serving as recipients and dictators along a "loop." In line with this paper's predictions and those of Hofmeier and Neuber (2019), Grech and Nax find less zero-giving in the interactive version of the dictator game than in the non-interactive one.

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 is probably unfamiliar to most potential voters. In contrast, egocentric norm adoption is grounded in people's intuition. It could explain why people sometimes resemble rule-utilitarians: 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), votes must exceed a certain threshold 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 as given the behavior of non-ethical voters and ethical voters who favor the opposite option.[34] However, one might still be puzzled why people who behave ethically in terms of incurring (individually useless) voting costs should disagree on 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 ethically demanded. Thus, a parsimonious behavioral principle explains prosocial behavior in turning out to vote and selfish behavior in terms of supported policies.

The models by Feddersen and Sandroni (2006a, 2006b) and Coate and Conlin (2004) both feature heterogeneous costs of voting. The rules that ethical voters adopt prescribe voting if and only if voting costs do not exceed a certain threshold value. That is because ethical voters aim at maximizing the utility of a group, and winning by an excessive margin would be wasteful. Thus, in the above models, heterogeneity enables coordination between voters who favor the same option. However, this model implication is at odds with experiments on public goods games that feature heterogeneity and find that heterogeneity reduces efficiency (see Fischbacher, Schudy, and Teyssier, 2014 and references therein).[35] 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 voting models and testing the resulting predictions would be a further exciting subject of future research.[36]

[34]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.

[35]Similarly, 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.

[36]For recent theoretical contributions, see Alger and Laslier (2020, 2021) and Grillo (2021).

Beyond voting and the examples in the introduction, many more 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 that unilateral action is pointless from a self-interested and strictly (act-) utilitarian perspective since costs are high and private returns (for a given country) are low. This insight 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, e.g., to the Marshall Islands, which are a small country in the Pacific Ocean that is part of Micronesia. However, the country is itself endangered by rising sea levels and has announced a plan for reducing carbon dioxide emissions to zero by 2050 (Malo, 2018). That a country with immense stakes takes bold steps against climate change, even when it has virtually no impact, is what egocentric norm adoption would predict. In this context, the behavioral phenomenon is also closely linked to setting an example (cf. Gächter et al., 2012; Gächter, Nosenzo, and Sefton, 2013). Indeed, Bicchieri et al. (2020) show that observing others breaching a norm erodes people's own propensity to comply with the norm, and others who obey a norm heighten compliance. This finding suggests that acting upon norms that one would want others to follow can be useful in the long term. It thereby provides a potential explanation for why the bias has been evolutionarily successful.

References

Acevedo, Melissa, and Joachim I. Krueger. 2005. "Evidential Reasoning in the Prisoner's Dilemma". *American Journal of Psychology* 118 (3): 431–457.

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

Alesina, Alberto, and Paola Giuliano. 2011. "Preferences for Redistribution". Chap. 4 in *Handbook of Social Economics*, ed. by Jess Benhabib, Alberto Bisin, and Matthew O. Jackson, vol. 1A, 93–131. North Holland.

Alger, Ingela, and Jean-François Laslier. 2021. *Homo moralis goes to the voting booth: a new theory of voter turnout*. Working Paper 1193. Toulouse: Toulouse School of Economics.

— . 2020. *Homo moralis goes to the voting booth: coordination and information aggregation*. Working Paper 1168. Toulouse, France: Toulouse School of Economics.

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.

Almås, Ingvild, Alexander W. Cappelen, Kjell G. Salvanes, Erik Sørensen, and Bertil Tungodden. 2017. "Fairness and family background". *Politics, Philosophy and Economics* 16 (2): 117–131.

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. 2020. "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?" *Experimental Economics*.

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).

Batson, C. Daniel, Bruce D. Duncan, Paula Ackerman, Terese Buckley, and Kimberly Birch. 1981. "Is Empathic Emotion a Source of Altruistic Motivation?" *Journal of Personality and Social Psychology* 40 (2): 290–302.

Batson, C. Daniel, Jim Fultz, and Patricia A. Schoenrade. 1987. "Distress and Empathy: Two Qualitatively Distinct Vicarious Emotions with Different Motivational Consequences". *Journal of Personality* 55 (1): 19–39.

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

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". *American Economic Review* 85 (1): 58–81.

Bicchieri, Christina, Eugen Dimant, and Silvia Sonderegger. 2020. *It's Not a Lie If You Believe the Norm Does Not Apply: Conditional Norm-Following with Strategic Beliefs*. CESifo Working Paper 8059. Munich, Germany: Munich Society for the Promotion of Economic Research – CESifo.

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.

Cappelen, Alexander W., Konow James, Erik Ø. Sørensen, and Bertil Tungodden. 2013. "Just Luck: An Experimental Study of Risk Taking and Fairness". *American Economic Review* 103 (4): 1398–1413.

Cassar, Lea, and Arnd H. Klein. 2019. "A matter of perspective: How failure shapes distributive preferences". *Management Science* 65 (11): 5050–5064.

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.

Chong, Alberto, and Mark Gradstein. 2008. "What determines foreign aid? The donors' perspective". *Journal of Development Economics* 87 (1): 1–13.

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.

Elster, Jon. 1998. "Emotions and Economic Theory". *Journal of Economic Literature* 36 (1): 47–74.

— . 1989. "Social Norms and Economic Theory". *Journal of Economic Perspectives* 3 (4): 99–117.

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, 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).

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study". *Journal of Political Economy* 127 (4): 1826–1863.

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.

Grech, Philip D., and Heinrich H. Nax. 2020. "Rational altruism? On preference estimation and dictator game experiments". *Games and Economic Behavior* 119:309–338.

Grillo, Alberto. 2021. *Ethical Voting in Heterogenous Groups*. Working Paper 34. Marseille, France: Aix-Marseille School of Economics.

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.*

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.

— . 2009. "Is fairness in the eye of the beholder? An impartial spectator analysis of justice". *Social Choice and Welfare* 33 (1): 101–127.

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

Krueger, Joachim I., and Melissa Acevedo. 2007. "Perceptions of self and other in the prisoner's dilemma: Outcome bias and evidential reasoning". *American Journal of Psychology* 120 (4): 593–618.

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.

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.

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.

— . 2015. "Kantian optimization: A microfoundation for cooperation". *Journal of Public Economics* 127:45–57.

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.

Singer, Tania, and Ernst Fehr. 2005. "The Neuroeconomics of Mind Reading and Empathy". *American Economic Review* 95 (2): 340–345.

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.

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.

Van Boven, Leaf, David Dunning, and George Loewenstein. 2000. "Egocentric Empathy Gaps Between Owners and Buyers: Misperceptions of the Endowment Effect". *Journal of Personality and Social Psychology* 79 (1): 66–76.

Van Boven, Leaf, George Loewenstein, David Dunning, and Loran F. Nordgren. 2013. "Changing Places: A Dual Judgment Model of Empathy Gaps in Emotional Perspective Taking". Chap. 3 in *Advances in Experimental Social Psychology*, ed. by Mark Zanna and James Olson, 48:117–171. Academic Press.

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.

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. The first order conditions for Equations 1 and 2 are as follows.

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

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 $\pi\alpha$ for α, with $\pi \in [0, 1]$. Thus, the likelihood of any set of values under the prior beliefs is

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

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.

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

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.

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

$$\tilde{\beta}_1 - \beta_1 = -\frac{(1-\pi)\,\alpha\, \mathit{ineff}\, \mathit{unfair}^2\, \mathit{pay}(\mathit{role}_{EF})}{D} \tag{A.2}$$

$$\tilde{\beta}_2 - \beta_2 = -\frac{(1-\pi)\,\alpha\, \mathit{ineff}^2\, \mathit{unfair}\, \mathit{pay}(\mathit{role}_{EQ})}{D} \tag{A.3}$$

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

Proof of Lemma 2. The first order conditions for Equations 3 and 4 are as follows.

$$\begin{aligned}
\tilde{\alpha}\, \mathit{Pay}'(\tilde{c}_{EF}, \mathit{role}_{EF}) - \tilde{\beta}_1\, \mathit{Ineff}'(\tilde{c}_{EF}) - \tilde{\gamma}\, \mathit{Inequal}'_{EF}(\tilde{c}_{EF}) = 0 \\
\tilde{\alpha}\, \mathit{Pay}'(\tilde{c}_{EQ}, \mathit{role}_{EQ}) - \tilde{\beta}_2\, \mathit{Unfair}'(\tilde{c}_{EQ}) - \tilde{\gamma}\, \mathit{Inequal}'_{EQ}(\tilde{c}_{EQ}) = 0
\end{aligned}$$

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 $\pi\alpha$ for α, with $\pi \in [0, 1]$. Thus, the

likelihood of any set of values under the prior beliefs is

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

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.

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

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

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

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

$$\tilde{\alpha} - \alpha = -\frac{(1-\pi)\,\alpha \left[\mathit{ineff}^2 \left(\mathit{unfair}^2 + \mathit{inequal}^2_{EQ}\right) + \mathit{unfair}^2\, \mathit{inequal}^2_{EF}\right]}{D} \tag{A.4}$$

$$\tilde{\beta}_1 - \beta_1 = -\frac{\begin{array}{r}(1-\pi)\,\alpha\, \mathit{ineff} \big[\left(\mathit{unfair}^2 + \mathit{inequal}^2_{EQ}\right) \mathit{pay}(\mathit{role}_{EF}) \\ - \mathit{inequal}_{EF}\, \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EQ}) \big]\end{array}}{D} \tag{A.5}$$

$$\tilde{\beta}_2 - \beta_2 = -\frac{\begin{array}{r}(1-\pi)\,\alpha\, \mathit{unfair} \big[\left(\mathit{ineff}^2 + \mathit{inequal}^2_{EF}\right) \mathit{pay}(\mathit{role}_{EQ}) \\ - \mathit{inequal}_{EF}\, \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EF}) \big]\end{array}}{D} \tag{A.6}$$

$$\tilde{\gamma} - \gamma = -\frac{\begin{array}{r}(1-\pi)\,\alpha \big(\mathit{ineff}^2\, \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EQ}) \\ + \mathit{unfair}^2\, \mathit{inequal}_{EF}\, \mathit{pay}(\mathit{role}_{EF}) \big)\end{array}}{D} \tag{A.7}$$

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 $\mathit{inequal}^2_{EQ}\, \mathit{pay}(\mathit{role}_{EF}) < \mathit{inequal}_{EF}\, \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EQ})$ implies that $\mathit{inequal}^2_{EF}\, \mathit{pay}(\mathit{role}_{EQ}) > \mathit{inequal}_{EF}\, \mathit{inequal}_{EQ}\, \mathit{pay}(\mathit{role}_{EF})$. 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

We conduct the following statistical hypothesis test:

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

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 in Equations 5 and 6, $1_A\big(r_i^{EF}\big)$ and $1_a\big(r_i^{EQ}\big)$ are statistically independent, since all combinations or roles appear with exactly the same frequencies in the experiment. Moreover, ϵ_i and η_i are each pairwise statistically independent of both $1_A\big(r_i^{EF}\big)$ and $1_a\big(r_i^{EQ}\big)$, 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 5 and 6 and retrieved the 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_δ^0 and t_ζ^0) 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.

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

The average of the separate two-sided p-values from the OLS regressions is thus an upper bound for p-value of the joint hypothesis test.

Appendix B Empirical Details

Table B.1: Sample Composition

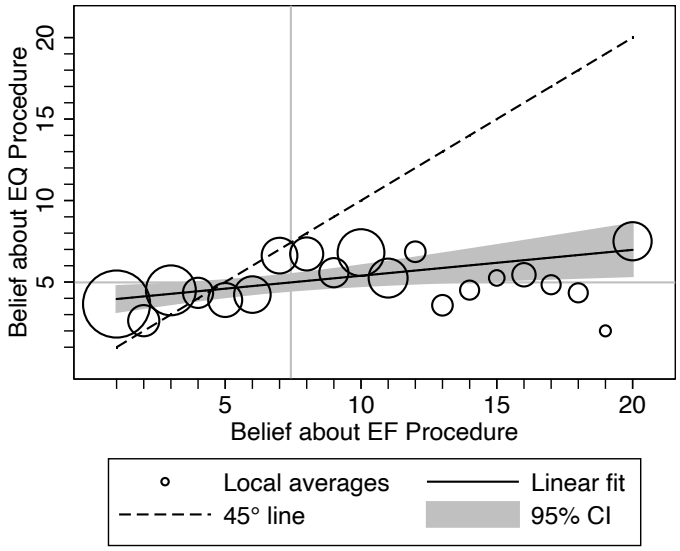
	Obs.	Mean	Median	Min.	Max.
Age	372	25.583	24	18	72
Female	369	0.599	1	0	1
University student	372	0.836	1	0	1
Income	372	741.185	600	0	3500
Log income	372	6.220	6.398595	0	8.160804

Notes: Log income is calculated as $\ln(income + 1)$

Table B.2: Beliefs

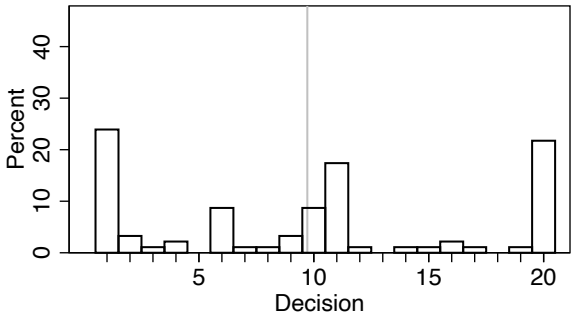
Dependent variable	*Belief about others' average decisions*			
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Role *A*	1.849***		1.852***	0.231
	(0.585)		(0.584)	(0.510)
Role *a*		2.102***	0.639	2.103***
		(0.510)	(0.584)	(0.510)
Constant	6.497***	3.925***	6.176***	3.809***
	(0.390)	(0.299)	(0.457)	(0.396)
Observations	372	372	372	372
R^2	0.026	0.044	0.029	0.045

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

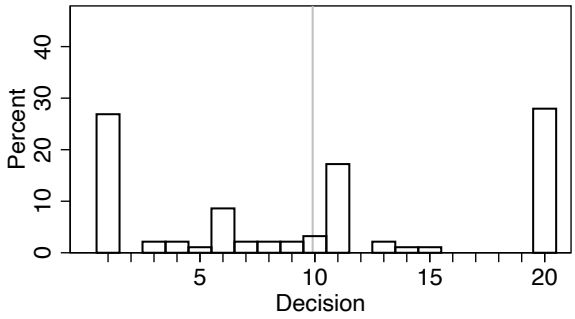


Notes: The figure groups subjects by their beliefs about others' average decisions for the EF Procedure. For each option on the horizontal axis, the figure plots the respective subjects' average belief about others' decisions for the EQ Procedure on the vertical axis. The sizes of circles correspond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

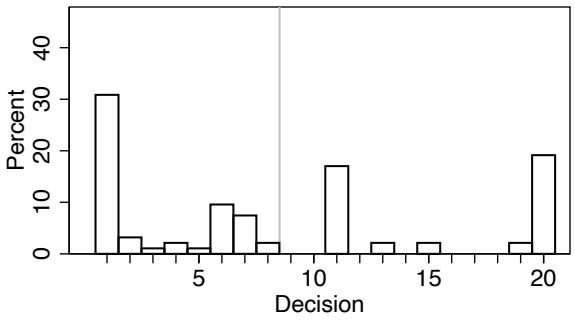
Figure B.1: Relationship Between the Two Predictions



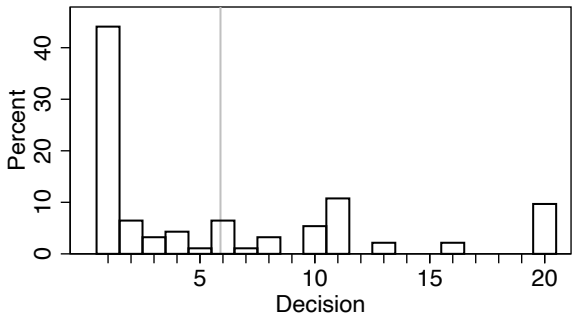
(a) Roles A and a



(b) Roles A and b



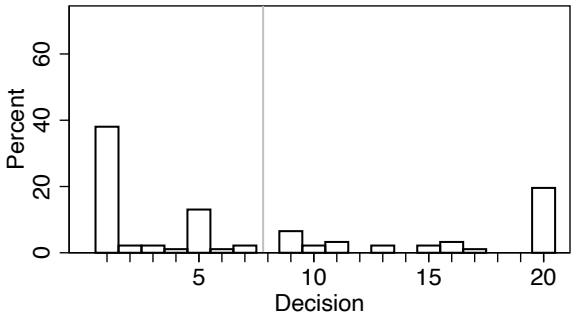
(c) Roles B and a



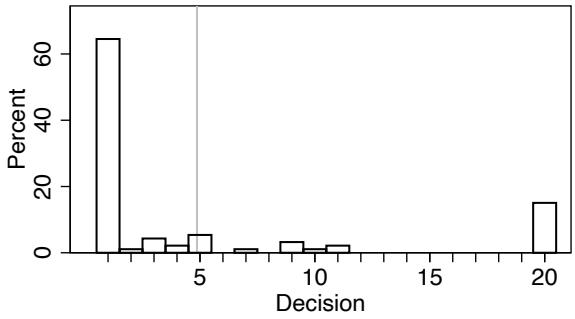
(d) Roles B and b

Notes: The gray lines indicate the respective average decisions.

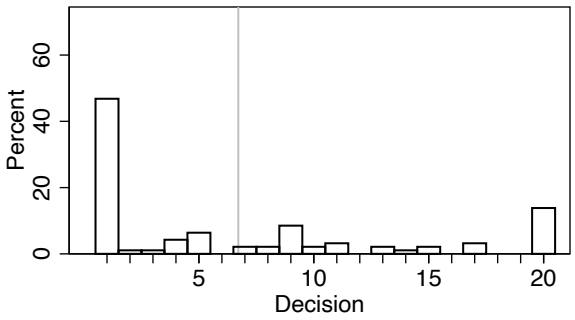
Figure B.2: Decisions for the EF Procedure by Combinations of Roles



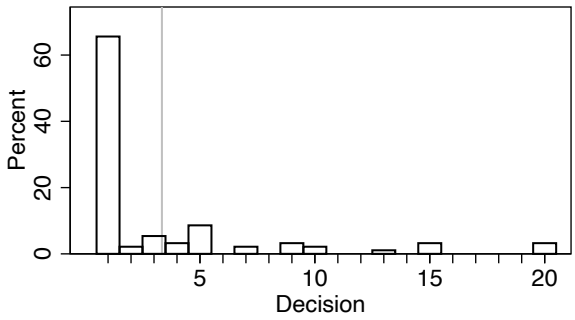
(a) Roles A and a



(b) Roles A and b



(c) Roles B and a



(d) Roles B and b

Notes: The gray lines indicate the respective average decisions.

Figure B.3: Decisions for the EQ Procedure by Combinations of Roles

Table B.3: Nominal Group Bias in Decisions

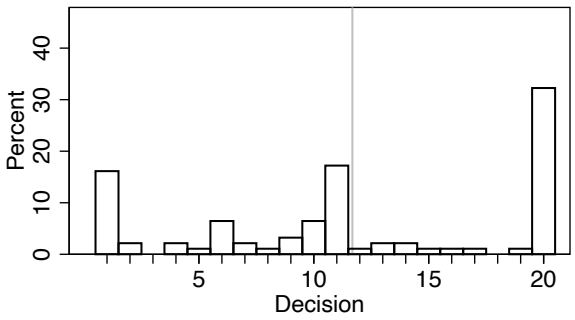
Dependent variable	*Decision*		*Belief*	
Procedure	*EF*	*EQ*	*EF*	*EQ*
	(1)	(2)	(3)	(4)
Same name is A	2.086***		0.958	
	(0.731)		(0.590)	
Same name is a		2.441***		1.715***
		(0.687)		(0.513)
Constant	7.454***	4.457***	6.935***	4.118***
	(0.503)	(0.437)	(0.417)	(0.300)
Observations	372	372	372	372
R^2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

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

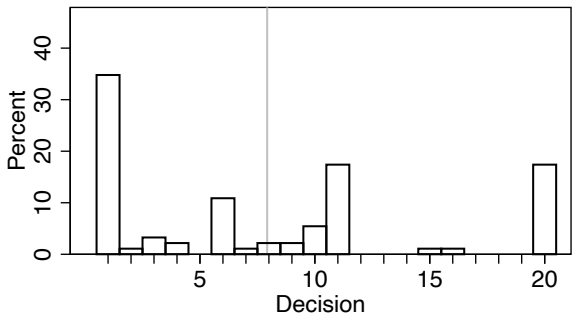
Table B.4: Nominal Group Bias in Decisions (with Roles)

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

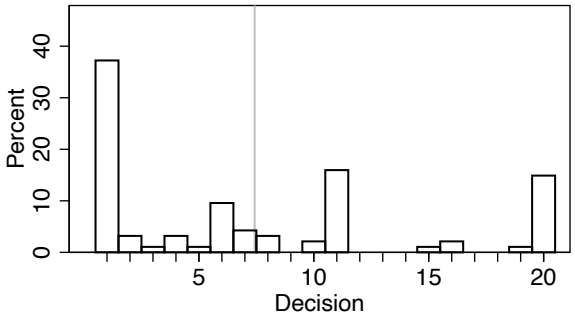
Notes: Heteroscedasticity-consistent standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.



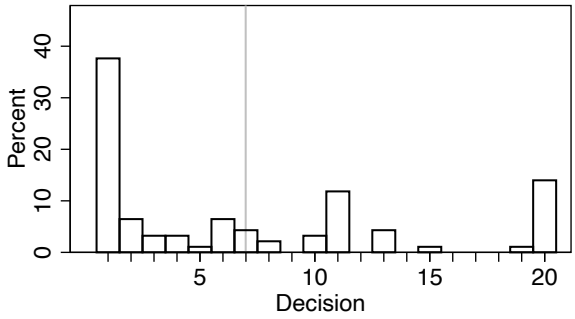
(a) Role A and player with same name is A



(b) Role A and player with same name is B



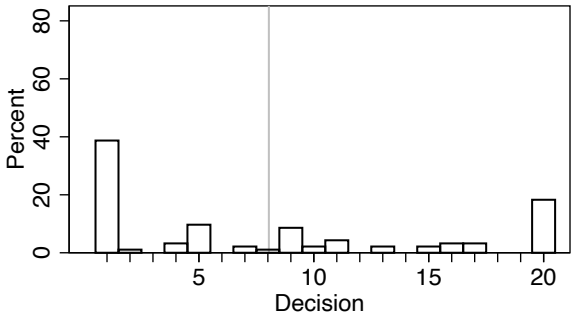
(c) Role B and player with same name is A



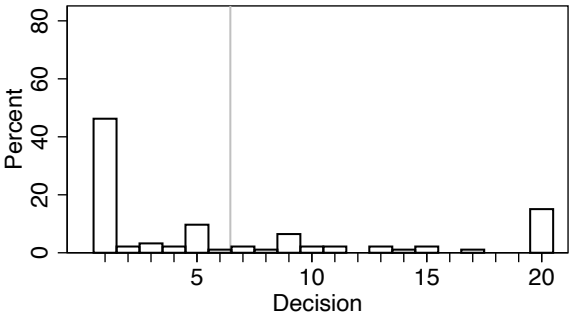
(d) Role B and player with same name is B

Notes: The gray lines indicate the respective average decisions.

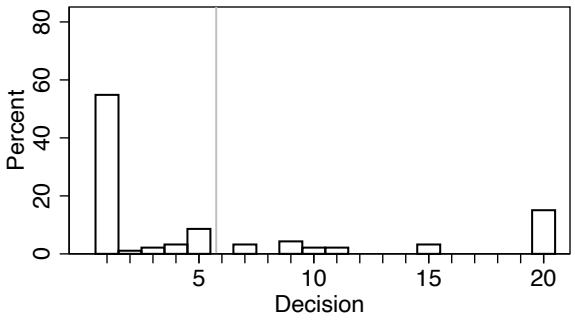
Figure B.4: Nominal Group Bias in the EF Procedure



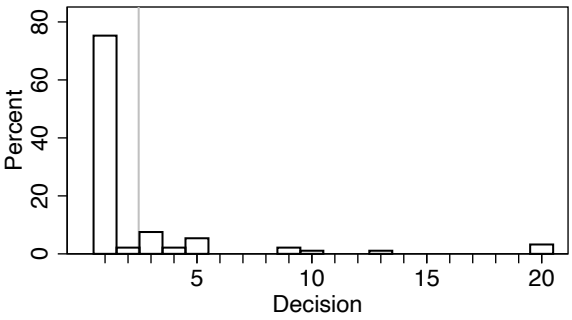
(a) Role a and player with same name is a



(b) Role a and player with same name is b



(c) Role b and player with same name is a



(d) Role b and player with same name is b

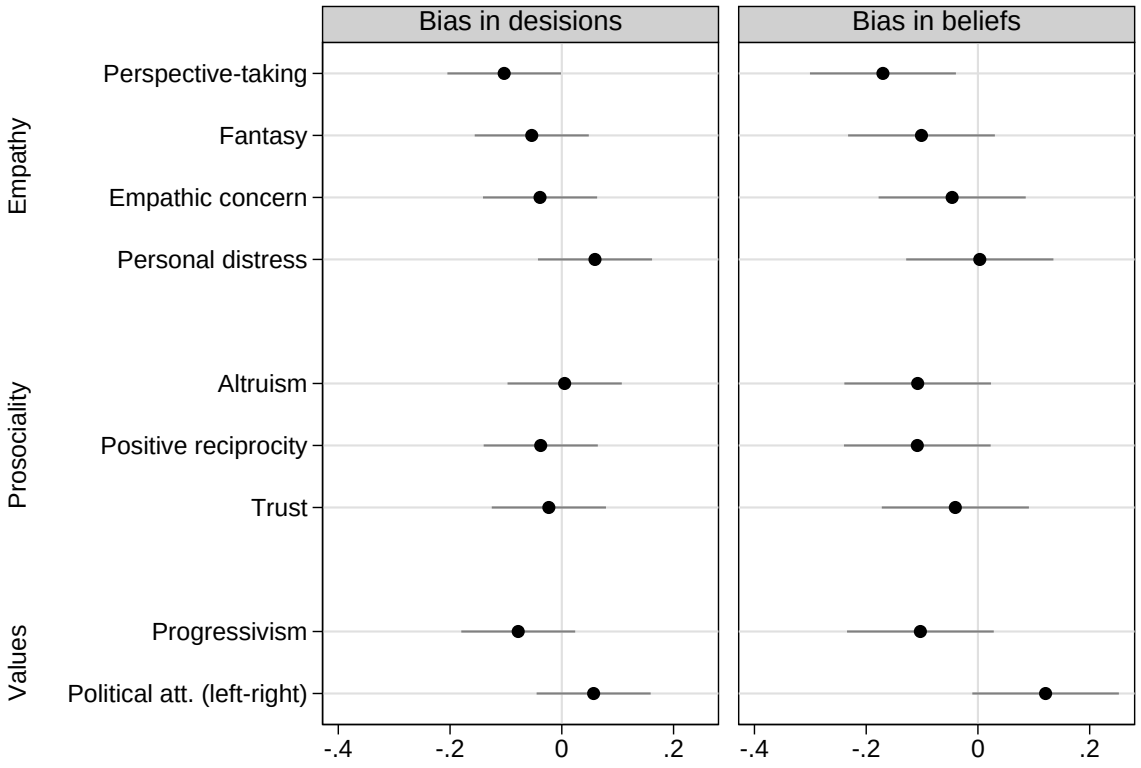
Notes: The gray lines indicate the respective average decisions.

Figure B.5: Nominal Group Bias in the EQ Procedure

Table B.5: 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

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



Notes: The figure shows the Pearson correlation coefficient for the ENA proxy introduced in Equation 7 and the respective survey measure. Gray bars indicate 95% confidence intervals.

Figure B.6: Correlations with the *ENA* Proxies (Full Sample)

Table B.6: Heterogeneity (showing controls)

Dependent variable	*ENA proxy*			
Domain	*Decisions*		*Beliefs*	
	(1)	(2)	(3)	(4)
Perspective-taking	-0.176**	-0.185**	-0.187**	-0.180*
	(0.0754)	(0.0754)	(0.0923)	(0.0936)
Fantasy	-0.0999	-0.0886	-0.132	-0.136
	(0.0760)	(0.0786)	(0.0934)	(0.0978)
Empathic concern	-0.00779	-0.0109	0.103	0.119
	(0.0783)	(0.0792)	(0.107)	(0.107)
Personal distress	0.248***	0.261***	0.131	0.123
	(0.0712)	(0.0719)	(0.0986)	(0.0991)
Altruism	-0.00313	-0.00408	-0.161*	-0.166*
	(0.0713)	(0.0718)	(0.0876)	(0.0868)
Positive reciprocity	0.0400	0.0473	-0.0349	-0.0196
	(0.0601)	(0.0611)	(0.0800)	(0.0802)
Trust	-0.111*	-0.104*	-0.0961	-0.0877
	(0.0606)	(0.0618)	(0.0761)	(0.0768)
Risk taking	0.155**	0.159**	0.244***	0.253***
	(0.0605)	(0.0623)	(0.0868)	(0.0880)
Patience	0.0224	0.0292	0.0300	0.0334
	(0.0627)	(0.0661)	(0.0869)	(0.0914)
Negative reciprocity	-0.126**	-0.138**	-0.0700	-0.0920
	(0.0600)	(0.0612)	(0.0722)	(0.0753)
Openness	0.0302	0.0362	-0.0310	-0.0320
	(0.0593)	(0.0623)	(0.0834)	(0.0855)
Conscientiousness	0.142**	0.145**	0.0994	0.0794
	(0.0619)	(0.0644)	(0.0804)	(0.0842)
Extraversion	-0.0272	-0.0265	0.0721	0.0751
	(0.0609)	(0.0614)	(0.0777)	(0.0795)
Agreeableness	0.0772	0.0839	0.115	0.112
	(0.0733)	(0.0761)	(0.0856)	(0.0849)
Neuroticism	-0.0919	-0.0808	0.00478	0.0157
	(0.0721)	(0.0730)	(0.103)	(0.102)
Female		-0.127		-0.0135
		(0.128)		(0.175)
Other gender		-0.378		0.703
		(0.343)		(0.704)
Age		0.113		-0.0489
		(0.232)		(0.404)
Age^2		-0.000352		-0.000117
		(0.000394)		(0.000811)
University student		-0.0448		-0.130
		(0.0762)		(0.0923)
Log income		0.0340		0.000219
		(0.0601)		(0.0688)
Personality controls	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes
Observations	312	312	312	312
R^2	0.120	0.130	0.093	0.106

Notes: 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 $\ln(income + 1)$.

Table B.7: Heterogeneity (Full Sample)

Dependent variable	*ENA proxy*			
Domain	*Decisions*		*Beliefs*	
	(1)	(2)	(3)	(4)
Perspective-taking	-0.151**	-0.158**	-0.200**	-0.197**
	(0.0706)	(0.0709)	(0.0822)	(0.0827)
Fantasy	-0.0378	-0.0337	-0.0535	-0.0615
	(0.0665)	(0.0681)	(0.0842)	(0.0868)
Empathic concern	0.0276	0.0278	0.140	0.159
	(0.0769)	(0.0776)	(0.0998)	(0.100)
Personal distress	0.166**	0.170**	0.102	0.0926
	(0.0685)	(0.0694)	(0.0871)	(0.0877)
Altruism	0.00246	0.00503	-0.145*	-0.146*
	(0.0642)	(0.0646)	(0.0789)	(0.0779)
Positive reciprocity	-0.0183	-0.0141	-0.104	-0.0935
	(0.0527)	(0.0538)	(0.0786)	(0.0789)
Trust	-0.0570	-0.0541	-0.0570	-0.0517
	(0.0556)	(0.0561)	(0.0727)	(0.0731)
Personality controls	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	Yes
Observations	372	372	372	372
R^2	0.078	0.083	0.071	0.079

Notes: 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 $\ln(income + 1)$.