Center for Mathematical Economics
Working Papers

666

May 2022

Fairness-based Altruism

Yves Breitmoser and Pauline Vorjohann

Center for Mathematical Economics (IMW)
Bielefeld University
Universitätsstraße 25
D-33615 Bielefeld · Germany

e-mail: imw@uni-bielefeld.de
uni-bielefeld.de/zwe/imw/research/working-papers
ISSN: 0931-6558

Unless otherwise noted, this work is licensed under a Creative Commons Attribution 4.0 International (CC BY) license. Further information:

https://creativecommons.org/licenses/by/4.0/deed.en

https://creativecommons.org/licenses/by/4.0/legalcode.en

Fairness-based altruism

Yves Breitmoser* Bielefeld University

Pauline Vorjohann University of Exeter

May 12, 2022

Abstract

Why do people give when asked, but prefer not to be asked, and even take when possible? We introduce a novel analytical framework that allows us to express context dependence and narrow bracketing axiomatically. We then derive the utility representation of distributive preferences additionally obeying standard axioms such as separability and scaling invariance. Such preferences admit a generalized prospect-theoretical utility representation reminiscent of fairness-based altruism. As in prospect theory, the underlying preferences are reference dependent and non-convex, which directly predicts the previously irreconcilable empirical evidence on giving, sorting, and taking. We test the model quantitatively on data from seminal experiments and observe significantly improved fit in relation to existing models, both in-sample and out-of-sample.

JEL codes: C91, D64, D03
Keywords: Social preferences, axiomatic foundation, robustness, giving, charitable donations

*We thank Niels Boissonnet, Friedel Bolle, Alexander Cappelen, Tore Ellingsen, Ernst Fehr, Urs Fischbacher, Susanne Goldlücke, Paul Heidhues, Steffen Huck, Dorothea Kübler, Matthias Lang, Sebastian Schweighofer-Kodritsch, Bertil Tungodden, Justin Valasek, Georg Weizsäcker, seminar participants in Berlin, Konstanz, Cologne, Düsseldorf, and Bergen as well as audiences in Munich (CESifo area conference), Kreuzlingen (theem 2019), Berlin (ESA 2018), Barcelona (IMEBESS 2017) and Lisbon (EEA/ESEM 2017) for many helpful comments. An earlier version of this paper was circulated with the title "Welfare-based altruism." Financial support of the DFG (project BR 4648/1 and CRC TRR 190) is greatly appreciated. Corresponding author: Yves Breitmoser. Address: Universitätsstr. 25, 33615 Bielefeld, Germany, email: yves.breitmoser@uni-bielefeld.de, phone: +49 521 106 5115.

1 Introduction

Altruism is widely defined as a concern for the well-being of others. At first sight, this definition appears to be self-explanatory, lending itself easily to economic modeling. Yet, any attempt at representing altruistic preferences by means of a utility function seems to prove the opposite. Andreoni and Miller (2002) showed that giving in the dictator game is well-captured by simple CES preferences, over the payoff pair of "dictator" and "recipient", which admits the intuitive interpretation that individual well-being is represented by CRRA utilities. Subsequent research showed, however, that no such utility function is compatible with giving in more realistic environments. For example, if we allow the recipient to have income of her own, giving is crowded out only imperfectly (Bolton and Katok, 1998), suggesting that warm glow may affect giving (Korenok et al., 2013). If we allow the dictator to take from the recipient's endowment, dictators largely stop giving, suggesting that context-independent altruism may not explain both giving and taking (List, 2007; Bardsley, 2008). Further, dictators that take tend to take as much as possible, suggesting non-convexity of preferences, and we observe asymmetries between giving and taking, suggesting that the warm glow of giving is weaker than the cold prickle of taking (Korenok et al., 2014). If we allow subjects to sort out of playing a dictator game (Dana et al., 2006), around half of them do so, including those who otherwise would give much to the recipient (Lazear et al., 2012), suggesting that we may need to augment altruism models by social-pressure motivation (DellaVigna et al., 2012). Overall, depending on the specific direction in which we extend the clinical dictator game, a different model of giving (and taking) seems to be required. This raises a fundamental question: Do the basic economic activities of taking, giving and sorting lend themselves to rigorous economic modelling at all? And if not, what does? Ultimately, any economic interaction essentially boils down to giving, taking and sorting.

In this paper, we present an axiomatic approach to the representation of preferences that allows us to directly address this potential impossibility. The main result is rather simple: when we follow Kahneman and Tversky (1979) and allow individual well-being to be S-shaped and reference dependent, while assuming that individuals care for the well-being of others, all the above phenomena are predicted immediately. Besides providing an intuitive and tractable model of the basic economic activities giving, taking and sorting, this preference representation connects distributive preferences to choice under risk—and perhaps most notably, it follows from an axiomatic analysis of choice. The axiomatic approach has two notable advantages in relation to existing work on modeling preferences. Existing work essentially seeks to construct models based on evidence from a given range of experiments. A model constructed in this way will be just one of many candidates compatible with the chosen evidence. This implies that a potential impossibility cannot be addressed. Furthermore, the chosen evidence tends to be context specific, implying that the resulting representation may not be portable to other contexts. In contrast, an axiomatic approach identifies all candidate models compatible with given behavioral regularities, including models that may not have been "invented" so far (such as our model), which allows us to address impossibility directly. Further, by shifting the focus from selected observations on taking and giving towards general behavioral regularities ("axioms") to build the foundation for modeling, we are potentially able to improve model portability substantially. Obviously, plausibility and generality of the underlying axioms are critical in this respect.

Our selection of axioms serves a simple objective. We aim to model a decision maker with preferences admitting an interpretation in relation to altruism (concern for the well-being of others) who is making decisions that are invariant to both scaling and translating outcome vectors, as both these invariances are compatible with existing evidence (reviewed below). The evidence notwithstanding, combining scaling and translation invariance in a single model is not straightforward. Previous results imply (see e.g. Skiadas, 2016) that if a preference ordering is invariant to scaling outcomes (multiplication by a common factor), then it cannot be invariant to translating outcomes (adding a common vector to all otucomes) unless the decision maker is indifferent between all options. Looking closer at the empirical evidence, however, the apparent translation invariance is usually observed in the form of narrow bracketing, i.e. independence of background income. We therefore introduce a generalized analytical framework that allows us to express inde-

pendence of background income (narrow bracketing) axiomatically. Our generalized framework distinguishes different decision problems to which we refer as contexts. This allows us to express narrow bracketing as a form of context dependence. Put this way, the implied translation invariance does not conflict with scaling invariance anymore. The relation to altruism theories follows from a separability axiom, besides axioms guaranteeing the existence of a continuous utility function. Finally, a novel "compensability" axiom allows us to connect preferences across different contexts by assuming that "compensatory incomes" are additive.

We find that preferences compatible with our axioms, i.e. primarily with scaling invariance, narrow bracketing, and compensability, always have a generalized prospect theoretical representation with reference points that are linear functions of observables (background income and default option). That is, the utility representation is a weighted mean of the individual well-beings of all agents involved ("altruism"), and the individual well-beings are represented by the reference-dependent value functions known from prospect theory. This provides a novel foundation for prospect theoretical utilities, indeed a first such foundation without assuming the existence of reference points. Furthermore, it gives a novel foundation for altruistic utilities and a first family of utility functions linking prospect theory and altruism. In that sense, our model is promising for a significant range of applications. Finally, the reference points are intuitively interpreted as fairness ideals or social norms, and this way, the identified preference representation relates to *fairness-based altruism* as discussed in the literature (Cappelen et al., 2007)[1], which we therefore adopt as name and implicitly characterize axiomatically.

We then investigate the extent to which this new representation of fairness-based altruism organizes the behavior observed in taking, giving and sorting games. To this end, we first derive and test a number of theoretical predictions for a reasonably comprehensive set of "purely distributive" decision tasks, i.e. the distribution of windfall gains without risk and uncertainty. This includes all decisions referenced in our research question as posed in the abstract, i.e. the standard dictator game, distribution games with non-trivial endowments, taking games, and sorting games. Importantly, observed choice even in these purely distributive decision tasks was previously suspected to be jointly incompatible with rational choice, which led to the aforementioned array of model extensions that render modeling of simple taking and giving intractable.

We show that a dictator's optimal transfer at an interior solution decreases in her own reference point while it increases in the recipient's reference point. This explains how a reallocation of initial endowments affects the optimal transfer, by shifting the players' reference points, and predicts imperfect crowding out. Another feature inherent in fairness-based altruism is that the resulting preferences are not convex, as individual welfares are S-shaped. Non-convexity directly explains that allowing the dictator to take from the recipient's initial endowment may result in "preference reversals", meaning that a dictator whose optimal choice in a game without the possibility to take is to transfer a positive amount to the recipient may switch to taking from the recipient once this is allowed (List, 2007; Bardsley, 2008). This explains why people would given when asked but also take when possible. In addition, the non-convexity predicts that when dictators "break the norm" and decide to take from the recipient's endowment, then they would tend to take it all, as observed by List (2007).

Relatedly, losses in relation to the reference point loom larger than gains, akin to loss aversion, explaining the asymmetries between giving and taking (Korenok et al., 2014). Fairness-based altruism also predicts the existence of "reluctant sharers", i.e. persons who transfer a positive amount to the recipient in a standard dictator game but choose a costly option to sort out of the game when given the chance (Dana et al., 2006; Lazear et al., 2012). Since the recipient never learns about the game if the dictator sorts out, her reference point is not adjusted to the dictator game environment and her welfare remains neutral. Once the dictator enters the game, the recipient is informed about the scope of the interaction and forms a reference point, which inflicts a negative externality on a fairness-based altruist. If the dictator believes the recipient would form a high reference point once informed, she is best off sorting out and leaving the recipient

[1]The difference of our version of fairness-based altruism to existing ones is that utilities in our case are S-shaped, while existing work studies quadratic utilities. Preferences represented by quadratic utilities are not scaling invariant in our sense.

uninformed—which explains why even "givers" often prefer not to be asked. It is worth noting that these predictions are explicit, i.e. the opposite results are ruled out by fairness-based altruism (in a sense made precise in Proposition 3), and that they all follow from a highly tractable model that directly generalizes representations of individual decision making.

To summarize, the aforementioned behavioral regularities (axioms), which formally represent a range of empirical results from meta-analyses and neuro-economics, directly predict the observations on taking, giving and sorting previously considered intractable, while incorporating many ideas from previous work: individuals are concerned with the well-being of others ("altruism"), individual well-beings are prospect theoretical (Kahneman and Tversky, 1979), and the reference points can be interpreted as expectations or social norms (as suggested Cappelen et al., 2007) that depend on the default outcome and on the minimum outcome (as conjectured by List, 2007, and Bardsley, 2008), all of which follows from an axiomatic analysis of rational choice.

In conjunction, these results provide a unified explanation for why people give when asked, but prefer not to be asked, and even take when possible. We acknowledge that the distributive problems we explicitly cover are, by far, not exhausting the full set of distributive situations discussed in the literature on social preferences. As indicated, we focus on one-shot distributive decisions made by single decision makers under certainty. This already makes for a large set of decision problems to cover theoretically and econometrically in a single paper. By focusing on non-strategic distributive problems, we avoid various confounds due to, for example, projection of preferences (as suspected in ultimatum games), coordination problems (as in public goods games), or non-rational expectations in strategic beliefs.[2] Furthermore, by excluding decision problems that involve uncertainty such as moral wiggle room experiments (Dana et al., 2007), we abstract from image concerns which alongside altruistic concerns are suspected to play a role for giving behavior. Finally, we do not cover distributive decisions which arise as a result of a larger history of play such as, for example, second-mover behavior in ultimatum games and trust games, where other factors like reciprocity (Cox et al., 2007; Dufwenberg and Kirchsteiger, 2004) and inequity aversion (Fehr and Schmidt, 1999) may kick in as additional motivations for behavior. Despite these restrictions of the present analysis, we are convinced that our analysis provides a valuable first step towards unifying the vast amount of evidence on decisions with implications for the well-being of others. Giving, taking, and sorting decisions in particular have previously been suspected to be incompatible with rational choice. We are aware, however, that most social interactions involve additional phenomena. Thanks to the generality of the axioms underlying our representation and the resulting relation to prospect theory, these phenomena will hopefully prove to be easier to organize once pure concerns for distribution are modeled consistently.[3]

After demonstrating how our model theoretically predicts the range of stylized facts on taking, sorting and giving, we evaluate whether fairness-based altruism indeed captures distributive decisions in these contexts quantitatively—in sample and in particular out of sample, which allows us to assess potential overfitting. To this end, we rely on data from controlled laboratory experiments. This data enables us to test our model very directly but, as reviewed below, the phenomena observed in the field are very similar.

It is worth noting that the concerns about overfitting, which we address here, apply equally to all models. In particular, they apply to all behavioral models generalizing the so-called standard models, regardless of whether they are models of choice, probability weighting, strategic beliefs, learning, or social preferences. Yet, outside the context of choice under risk (Harless et al., 1994; Wilcox, 2008; Hey et al., 2010), analyses quantitatively testing robustness are comparably rare.[4] This has been taken as a suggestion that many behavioral models may lack robustness. We theoretically and econometrically demonstrate the opposite for our model of fairness-based altruism, ruling out overfitting as a concern.

[2]For a discussion of the mentioned confounds, we refer the interested reader to Blanco et al. (2011).

[3]For example, image concerns are straightforward to add once altruism is modeled consistently, and distributive choice under risk may be implied by the relation to prospect theory.

[4]The short list of exceptions that we are aware of comprises analyses of learning (Camerer and Ho, 1999), strategic choice in normal-form games (Camerer et al., 2004), stochastic choice in dictator games (Breitmoser, 2013, 2017), bargaining preferences (De Bruyn and Bolton, 2008), and most recently, social preferences (Bruhin et al., 2018).

We first estimate the distributions of individual reference points in the four types of distribution games representing the corner stones of the current debate: standard dictator games, distribution games with generalized endowments, taking games, and sorting games. The estimated reference points are surprisingly consistent and we identify three clusters resembling the non-givers, altruistic givers, and social-pressure givers observed by DellaVigna et al. (2012) in charitable fundraising. Implicitly, this clarifies how the diversity of altruism types is captured in a formally uniform manner by fairness-based altruism. Further, it shows that adjustments of reference points do not drive model adequacy across conditions.

In a second step, we re-analyze behavior across a set of nine well-known laboratory experiments on distribution games, comprising 83 choice conditions and around 6500 decisions from 981 subjects. Besides improving in-sample fit, we find that compared to the standard CES model of altruism, predictions improve substantially by allowing for fairness-based altruism. Notably, this holds consistently across all combinations of in- and out-of-sample conditions. We then examine two alternative approaches of extending the standard CES model that are proposed in the literature, capturing either warm glow and cold prickle, or envy and guilt. They both fail to improve on CES altruism out-of-sample. Our results confirm the general suspicion that achieving out-of-sample robustness is indeed challenging when modeling distributive concerns. Indeed, this was our initial reason to pursue an axiomatic approach based on general behavioral traits. Overall, our econometric analysis shows that the identified generalization of prospect-theoretic utilities towards fairness-based altruism is a promising approach for modeling distributive concerns. Furthermore, by unifying social preferences and risk preferences, our model can provide a useful framework for future work seeking to capture distributive concerns in strategic interactions and under uncertainty.

The paper is organized as follows. After a brief review of the related literature, Section 2 gives an overview of recent evidence on decisions in distribution problems. Section 3 provides our representation result (Proposition 1) and discusses its relation to the literature in detail. Section 4 analyzes the implications for giving theoretically in relation to the stylized facts summarized in Section 2 (Propositions 2 and 3). Section 5 evaluates fairness-based altruism econometrically using a range of in-sample and out-of-sample analyses. Section 6 concludes. The appendix contains relegated definitions, proofs, and robustness checks.

1.1 Related literature

Similar to us, Becker (1974) treats altruism as a concern for the utility of others. His representation, however, yields a linear equation system that can be solved to represent altruism as a pure concern for payoffs of others. The resulting differences to standard models in games of complete information are negligible (Kritikos and Bolle, 2005). Even earlier, Edgeworth considered general models of altruism that contain ours as a special case. In these models, altruism is a concern for "internal utilities" (Dufwenberg et al., 2011) of others, without the particular prospect-theoretical formulation that we show our axioms to imply. Therefore, our results relate back to this classical idea, based on which, most notably, Dufwenberg et al. (2011) show that agents behave *as-if-classical* in markets, in the sense that their demand functions depend only on own income and prices. This preserves the applicability of standard techniques and allows them to demonstrate that the Second Welfare Theorem continues to hold.

Cox et al. (2016) present an axiomatic approach to modeling moral costs in distributive decisions that arise from dictators not satisfying what they call "moral reference points". They assume that these reference points are determined by players' minimal payoffs and their initial endowments, which is implied by our results. Positing an axiom that assures a form of monotonicity of the dictator's choices with respect to reference points, they show that a modified version of the contraction axiom also known as property α (Chernoff, 1954; Sen, 1971) is met. Importantly, this modification allows for non-convexity of preferences which helps reconcile experimental evidence from classical dictator and taking games. Their paper is a great complement to our study. While we mainly focus on the specific functional form of a utility function representing distributional preferences, showing that it is indeed reference dependent, Cox et al. (2016) are mainly concerned

with the specific form of reference points in distributive decisions, not taking a stance on how exactly these reference points may enter a dictator's utility function.

The fairness-based altruist's concern for the recipient's reference-dependent well-being as implied by our model is also intuitively related to the concept of guilt aversion (Charness and Dufwenberg, 2006). Guilt aversion posits that decision makers experience guilt when they do not fulfill others' expectations of their behavior. Similarly, the fairness-based altruist suffers a utility loss if she does not fulfill another player's reference point. The main difference between our model and the idea of guilt aversion is that reference points capture a broader concept of what a player "expects" from an interaction than pure beliefs about another player's behavior. In particular, models of guilt aversion commonly rely on psychological game theory in the sense that beliefs are assumed to be correct in equilibrium (Geanakoplos et al., 1989). This reliance implies that they have no bite in the distributive decisions under certainty covered here.

Our axiomatic analysis establishes a somewhat unsuspected interdependence of concepts as diverse as prospect theory, narrow bracketing, altruism, social appropriateness (discussed below), and reference dependence—besides predicting a range of behavioral puzzles that survived for about 20 years of experimental research. This underlines the adequacy of axiomatic analyses towards understanding social preferences. Further, our results imply that decision makers are utilitarists (for recent discussions, see Fleurbaey and Maniquet, 2011, and Piacquadio, 2017) but in a manner that was predicted by Rawls: rational agents "do not take an interest in one another's interests" (Rawls, 1971, p. 13). That is, agents are concerned with the well-being of others in the way that these others would perceive it in one-person decision problems, but they are not concerned with their altruism or envy, for example. This in turn provides a normative argument for "preference laundering" (Goodin, 1986) in behavioral analyses of social welfare, i.e. for the neglect of emotions such as altruism or envy in welfare analyses.

The work of Rawls also bridges our findings to "social appropriateness" of distributions as discussed by Krupka and Weber (2013). Their results suggest that behavior may be norm-guided rather than payoff or well-being concerned, casting general doubt on the applicability of models (such as ours) proposed in the existing literature. In Appendix A, we show that the measure of "social appropriateness" they elicit via coordination games strongly correlates with the Rawlsian notion of social welfare as implied by our out-of-sample predictions of each player's individual welfare (it is an affine transformation of the minimum of these individual welfares). That is, we show that social appropriateness has a simple and intuitive Rawlsian foundation in individual well-being—which we interpret to lend further credibility to both, fairness-based altruism and social appropriateness, as dual approaches towards analyzing behavior.

Finally, by generalizing prospect-theoretical utility, fairness-based altruism addresses a number of practical concerns in the literature, such as providing a unified framework for measuring robustness and heterogeneity of preferences across populations and decision problems (Falk et al., 2018), providing a normatively founded framework for measuring reference points across interactions, thereby facilitating a solution to the long-lasting debate on whether and when reference points reflect a status quo (Kahneman et al., 1991), expectations (Kőszegi and Rabin, 2006), or others' payoffs (Fehr and Schmidt, 1999), and providing a general framework for structural analyses of charitable giving (DellaVigna et al., 2012; Huck et al., 2015).

2 Experimental evidence on giving

We are analyzing a variety of distribution problems under complete information, each of which is more or less closely related to the classic dictator game. In each game, there are two players, the dictator and the recipient. Player 1 (dictator) is endowed with B_1 tokens and player 2 (recipient) is endowed with B_2 tokens. Player 1 can choose $p_1 \in P_1 \subset \mathbb{R}$, inducing a payoff of p_1 for herself and a payoff of $p_2(p_1) = t(B_1 + B_2 - p_1)$ for player 2. We refer to $t > 0$ as transfer rate, to $B = B_1 + B_2$ as budget, and to $B_1 - p_1$ as transfer from the dictator to the recipient.

Definition 1 (Distribution game)**.** A distribution game Γ is defined by the tuple $\langle B_1, B_2, P_1, t \rangle$. The

following variants will be distinguished.

- *Standard dictator game:* $B_1 > 0, B_2 = 0, P_1 \subseteq [0, B_1]$
- *Generalized endowments:* $B_1 \geq 0, B_2 > 0, P_1 \subseteq [0, B_1]$
- *Taking game:* $B_1 \geq 0, B_2 > 0, P_1 \subseteq [0, B_1 + B_2]$
- *Sorting game:* $B_1 > 0, B_2 = 0, P_1 \subseteq \{[0, B_1], \tilde{p}_1\}$ where $\tilde{p}_1$ is an outside option for player 1 inducing payoffs $(\tilde{p}_1, 0)$, with $\tilde{p}_1 \leq B_1$, and implying that 2 is not informed about 1's choice or the rules of the game.

Table 1 provides an overview of the behavior observed in these distribution problems. Following the early work of Kahneman et al. (1986) and for example Hoffman et al. (1996), comprehensive analyses of behavior in the standard dictator game are presented in Andreoni and Miller (2002) and Fisman et al. (2007). The authors show that the average share of the budget transferred by dictators varies between 20% and 30%, there is an accumulation of transfers at zero and at the payoff-equalizing option, and there is considerable heterogeneity in transfers between subjects. Furthermore, varying budget sets B and transfer rates t, observed transfers to a large extent satisfy the generalized axiom of revealed preference, implying that dictator behavior is consistent with well-behaved preference orderings. As a candidate for a utility function representing these preferences, Andreoni and Miller (2002) proposed the CES model of altruism, which, using the formulation of Cox et al. (2007), is given by

$$u(\pi) = \pi_1^\beta/\beta + \alpha\, \pi_2^\beta/\beta \qquad \text{(CES altruism)}$$

with $\alpha, \beta \in \mathbb{R}$. Here, α represents the degree of altruism, $\beta = 1$ implies efficiency concerns, $\beta \to 0$ yields Cobb-Douglas utilities, and $\beta \to -\infty$ implies equity concerns (Leontief preferences).

Comparative statics in t In a meta-analysis of about 100 experiments, Engel (2011) shows that dictators' transfers increase in the transfer rate, i.e. as transfers become more efficient. This has been observed earlier by Andreoni and Miller (2002) but, for example, not by Fisman et al. (2007). The individual level analyses of Andreoni and Miller (2002) and Fisman et al. (2007) suggest that this inconsistency may be due to differences in subject heterogeneity. In both studies, the majority of subjects act consistently with CES altruism and can be weakly categorized into three standard cases of this utility function, namely selfish, perfect substitutes, and Leontief. Perfectly selfish preferences imply no reaction to changes in the transfer rate, but dictators increase transfers after increases of t if they consider the payoffs to be imperfect substitutes ($\beta > 0$), and they decrease transfers if they consider payoffs to be imperfect complements ($\beta < 0$).

Taking options reduce giving at the extensive and intensive margin Holding initial endowments constant, convexity of preferences implies that the extension of the dictator's option set to negative transfers does not affect the choice of a dictator unless she chooses the boundary solution of giving nothing in a standard dictator game. This prediction is implied by most models of giving, including CES altruism for $\beta < 1$, but falsified by a strand of studies on so-called taking games (List, 2007; Bardsley, 2008). Both List and Bardsley found that introducing options to take reduces the share of dictators who give positive amounts, though not always significantly. Furthermore, it reduces average amounts given by those who do give positive amounts and leads to substantive accumulation at the most selfish option. Korenok et al. (2014) confirm these results. List (2007) and Cappelen et al. (2013b) obtain related results on real-effort versions of taking games. List (2007) and Bardsley (2008) interpret the observed patterns in taking games as an indication that choice is menu dependent and, for example, Korenok et al. (2014) argue that taking might induce cold prickle in the sense of Andreoni (1995). In contrast, we argue that the initial assumption of convexity may be violated, as known, for example, from choice under risk.

Table 1: Stylized facts about distribution games

Comparative statics in t	The transfer can be either constant, increasing, or decreasing in the transfer rate.
Taking options reduce giving at the extensive and intensive margin	Holding endowments constant, extending the choice set of the dictator to the taking domain transforms some initial givers into takers and reduces average amounts given.
Incomplete crowding out	Reallocating endowment from the dictator to the recipient while holding the overall budget constant leads to a less than one-to-one reduction in the dictator's transfer.
Reluctant sharers	A substantial share of givers in the standard dictator game choose to sort out of the game when given the opportunity.
Outside option attractiveness	As the outside option becomes less attractive, fewer dictators sort out of the game. Nonsharers sort back in first followed by the least generous sharers and successively more and more generous sharers.

Incomplete crowding out Reference independence of social preferences, as in CES altruism, implies the so-called *crowding out hypothesis* (Bolton and Katok, 1998): lump-sum transfers from dictator to recipient result in a dollar-for-dollar reduction in voluntary giving. The experimental results on dictator games with generalized endowments unanimously falsify this prediction. In both lab and field experiments, dictators reduce their transfers in response to reallocations of endowments to the recipient, but the observed reduction is significantly lower than predicted, a phenomenon referred to as incomplete crowding out (Bolton and Katok, 1998; Eckel et al., 2005; Korenok et al., 2012, 2013). These findings extend to the domain of taking games (Korenok et al., 2014) and to interactions where the budgets are not windfall but generated through either investment games or real effort tasks (Konow, 2000; Cappelen et al., 2007, 2010, 2013a; Almås et al., 2010; Ruffle, 1998; Oxoby and Spraggon, 2008; Jakiela, 2011, 2015). The evidence on dictator games with endowments generated in real effort tasks further suggests that the origin of initial endowments affects dictator behavior. Compared to a standard dictator game with windfall budget, the change to real effort budgets earned by the dictators leads to a drastic reduction in the proportion of nonzero transfers (Cherry, 2001; Cherry et al., 2002; Cherry and Shogren, 2008; Oxoby and Spraggon, 2008; Jakiela, 2011, 2015; Hoffman et al., 1994). Cappelen et al. (2007) relate the observed endowment effects to social norms and, for example, Korenok et al. (2013) interpret the endowment effects as a sign that warm glow in the sense of Andreoni (1995) affects giving. Outside the literature on social preferences, endowment effects are mostly related to reference dependence of preferences (Kahneman et al., 1991; Tversky and Kahneman, 1991), which in turn will be predicted by our representation result.

Reluctant sharers & outside option attractiveness Turning to sorting games, convexity of preferences implies that a dictator cannot be strictly better off by opting out than by staying in. For, the dictator game offers a budget that is at least as high as the outside option. Convexity also implies that no dictator who transfers a positive amount in the dictator game will opt out, since for such a dictator the outside option must be strictly worse than the allocation she chose in the dictator game. Falsifying this prediction, Dana et al. (2006), Broberg et al. (2007), and Lazear et al. (2012) find that a substantive share $(20-60\%)$ of their subjects in sorting games can be classified as reluctant sharers, i.e. as dictators who transfer a positive amount in the standard dictator game but given the opportunity rather opt out. As a result, the average amount shared significantly decreases when a sorting option is added to the standard dictator game. Lazear et al. (2012) also find that (i) making the outside option less attractive while holding the dictator game budget constant does reduce the number of dictators who opt out, but (ii) it also reduces the average amount shared. For,

mostly nonsharers and reluctant sharers who share less generously in the dictator game reenter first when opting out becomes less attractive. DellaVigna et al. (2012) and Andreoni et al. (2017) obtain similar results in field experiments on charitable giving. Related to that, Cappelen et al. (2017) observe a close interaction between the information the recipient receives about the origin of her payment and the transfers made by the dictators (in standard dictator games). There are again multiple proposals for capturing sorting theoretically. DellaVigna et al. (2012) suggest allowing for an aversion to "saying no" when asked about donations, which however does not capture the comparative statics observed by Lazear et al. (2012). Andreoni and Bernheim (2009) and Ariely et al. (2009), in contrast, propose capturing reluctancy by including image concerns. As indicated above, the falsified predictions for sorting games are closely related to convexity. Thus, the non-convexity of preferences implied by our model allows us to directly capture reluctancy and sorting decisions.

3 Fairness-based altruism: Axiomatic foundation

In this section, we introduce the behavioral assumptions characterizing the decision makers we have in mind and then derive the families of utility functions representing their preferences. These utility functions will be applied subsequently to analyze distributive decisions. Thus, we seek to test whether the axiomatic approach indeed identifies a family of utility functions with the potential of providing a unified explanation of the seemingly inconsistent giving, taking and sorting observations discussed above.

Wherever possible, our analysis is based on assumptions that are comparably well-accepted in related work, for example on choice under risk. We do, however, extend previous work in several important ways, most notably by distinguishing contexts to express narrow bracketing and context dependence alongside scaling invariance. This provides a novel foundation for reference dependence without explicitly assuming the existence of reference points, as we discuss in detail below.

We are not aware of directly comparable work on the foundation of interdependent preferences. There do, however, exist axiomatic foundations of inequity aversion, e.g. Rohde (2010) and Saito (2013), which provide insightful foundations for the widely-used model of Fehr and Schmidt (1999). The difference is that these approaches explicitly use inequity-aversion axioms to establish a foundation for this particular model—their objective is not to identify a general set of candidate models based on axioms not directly related to specific preconceptions of preference interdependence, which is what we attempt here.

3.1 Theoretical framework

Decision maker DM has to choose an option $x \in X$. Each option induces an n-dimensional outcome vector, as described by the outcome function $\pi : X \to \mathbb{R}^n$, with $n \geq 3$.[5] The reader may think of X as a convex subset of $\mathbb{R}^n$. We refer to π as an outcome function. While nothing in our theoretical analysis is specific to preferences over payoff profiles, monetary outcomes are a standard application. There exists a specific option "0"$\in X$ that we use to denote the default outcome $\pi(0)$, which will help us to discuss how preferences may depend on the default. Note, however, that our analysis is not specific to $\pi(0)$ being a default in a physical sense. This implies that $\pi(0)$ might represent any other benchmark of interest to the analyst. Generalizations towards decision making in the presence of several such benchmark options are straightforward, at the expense of additional notation.

[5] Assuming $n \geq 3$ simplifies some of the statements made below regarding existence of an additively separable utility representation. It is not crucial for the main result. If there was only one essential dimension, existence of an additively separable representation would be trivial, and if there were exactly two essential dimensions in the outcome vector, then existence of an additively separable representation would be ensured by additionally assuming the hexagon condition of Wakker (1989, p. 47).

We use Π to denote the set of outcome functions for which our assumptions shall hold. The reader may think of Π as the set of all functions $\pi : X \to \mathbb{R}^n$, but the formal requirements are provided below. The image of any $\pi \in \Pi$ is denoted as $\pi[X] = \{\pi(x)|x \in X\}$. By $\min\pi$ we refer to the component-wise minimum of π, i.e. the minimum $\min\pi = \{\min_x \pi_i(x)\}_{i\leq n}$ of π in all dimensions. We call $\min\pi$ *background income* vector. In a dictator game, $\min\pi$ is the vector of minimum incomes of all players. In some applications, the background income can also be interpreted as status quo, and it may equate with the default. We shall not restrict how background income, status quo, and default relate to each other.

DM has a preference ordering over the outcomes induced by options $x \in X$ that may depend on π in non-trivial ways. For example, take any two π and π' and two pairs of options (x,y) and (x',y') such that the associated outcomes are pairwise identical: $\pi(x) = \pi'(x')$ and $\pi(y) = \pi'(y')$. We allow for the possibility that DM prefers x over y in context π but y' over x' in context π'. This could, for example, arise because outcomes below reference points are ordered differently than outcomes above reference points while reference points may change when the set of feasible outcomes or the default changes. With our notation, we explicitly allow for such dependence on the contextual information contained in π. We shall therefore say that π characterizes the "context" of the decision, or simply that π is the context. This distinction of contexts is novel in relation to the literature and will allow us to state assumptions about responses to changes in background income, default, or concurrent tasks, as discussed below. Our decision to maintain notational convenience by equating "context" with π restricts the types of context dependence we can express. Specifically, we can express preferences that depend on the range of outcomes, the background income, the default, and other measures of the outcome function, but we cannot express dependence on prior events that contributed to the outcome function (say, whether luck, effort or manna from heaven has generated the default). Generalizations towards the latter are straightforward by extending the notation of contexts to also include the source of income and strengthening the limited context dependence ("narrow bracketing") axiom introduced below in order to acknowledge this extended understanding of contexts. It would, however, require additional notation that we seek to avoid here.

The context-dependent preference ordering on outcomes $\pi[X]$ is denoted as $\succsim_\pi$, with $\pi(x) \succsim_\pi \pi(y)$ indicating that outcome $\pi(x)$ is weakly preferred to outcome $\pi(y)$ in context π. Given π and $\succsim_\pi$, DM's preference relation R over option set X is straightforwardly defined as xRy if and only if $\pi(x) \succsim_\pi \pi(y)$, for all $x,y \in X$. As usual, the strict preference $\pi(x) \succ_\pi \pi(y)$ indicates $\pi(x) \succsim_\pi \pi(y)$ and $\pi(y) \not\succsim_\pi \pi(x)$. Finally, we use d to denote distance measures over the set of options X and sets of outcomes $\pi[X]$.

Given this notation, we impose four pieces of structure on the set of decision tasks (outcome functions) that we analyze. To provide intuition, we illustrate these assumptions in relation to n-player dictator game experiments (1 dictator, $n-1$ recipients) where X is the set of options available to the dictator and π is the mapping from options to payoff profiles. In such experiments, the outcome function π accounts for show-up fees, initial endowments of the agents, and conversion rates from experimental points towards local currency, all of which may be asymmetric. Our first assumption *context richness* requires that the experimenter can change initial endowments and conversion rates arbitrarily, and that she can combine any two decision problems π and π' towards a joint decision problem $\pi + \pi'$ (i.e. that a dictator decision has implications according to two outcome functions). Second, *default richness* requires that the experimenter can set the default arbitrarily, i.e. the default can be any option. Third, *option richness* requires that the set of options can be thought of as generated by a budget constraint where (i) the budget does not need to be exhausted, as in Andreoni and Miller (2002) for example, and (ii) it is possible for the experimenter to design an outcome function with an option where all but one of the agents are allocated nothing (trivially satisfied in typical dictator games). Finally, *essentialness* requires that there are no redundant dimensions of the outcome vector from DM's perspective, i.e. DM does not ignore any of the dimensions, which is a necessary condition for uniqueness of the utility representation in all dimensions. Formally, our conditions are defined as follows.[6]

[6]Slightly abusing notation, we identify all $c \in \mathbb{R}^n$ with constant functions so the addition of functions and constants is

Assumption 1 (Framework)**.**

1. *Context richness:* There exists a non-degenerate interval $\Lambda \subseteq \mathbb{R}_+$ with positive length such that $c + \lambda\pi \in \Pi$ for all $(c, \lambda) \in \mathbb{R}^n \times \Lambda$ and all $\pi \in \Pi$. Further, $\pi + \pi' \in \Pi$ for all $\pi, \pi' \in \Pi$.

2. *Default richness:* For each $\pi \in \Pi$ and each $x \in X$, there exists $\pi' \in \Pi : \min \pi' = \min \pi$ where $\pi'(0) = \pi(x)$.

3. *Option richness:* $\pi[X] - \min \pi$ is a non-degenerate cone in $\mathbb{R}^n$, i.e. for all $x \in X$ and all $\lambda \in [0,1]$, there exists $x' \in X$ such that $\pi(x') = \min \pi + \lambda(\pi(x) - \min \pi)$. Further, $\forall i \neq n$ there exists $\pi \in \Pi$ and $x \in X$ such that $\pi_i(x) > \pi_j(x) = 0$ for all j.

4. *Essentialness:* All $n \geq 3$ dimensions are essential, i.e. for all $i \leq n$ and each $\pi \in \Pi$, there exist $p, p' \in \pi[X]$ such that $p \succ_\pi p'$ with $p_{-i} = p'_{-i}$.

3.2 Axioms and representation result

We analyze the interplay of two sets of axioms. The first four axioms capture behavior in any given context and the remaining four axioms capture reactions to changes in context. The axioms are formally defined below but let us first provide a more intuitive description of the axioms in order to discuss the extent to which they represent reasonable descriptions of behavior. Axioms (1) and (2) require that $\succsim_\pi$ is a continuous weak order, implying that it can be represented by a continuous utility function, which we need in order to discuss DM's utility function.

Separability (Axiom (3)) is also known as "independence of equal coordinates" (Wakker, 1989, p. 30). It implies additive separability of the utility function, defined formally below. In relation to a DM in dictator games, separability implies that we can think of DM as being concerned with the welfare (or payoff) of others, assuming that each individual welfare is a function of the respective individual payoff. In this sense, separability represents an "altruism axiom" in our analysis. Relatedly, additive separability obtains in most utility representations discussed in the literature on altruistic giving, such as CES altruism (Andreoni and Miller, 2002), efficiency concerns (Charness and Rabin, 2002), and impure altruism (Andreoni, 1990; Korenok et al., 2013). In turn, non-separable preferences are typically used to model phenomena not related to altruism, such as inequity aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). More generally, separability closely relates to a broad range of standard assumptions: independence axioms in choice under risk (Wakker and Zank, 2002) or choice under uncertainty (Skiadas, 2013), "independence of irrelevant alternatives" in stochastic choice (Luce, 1959), and separability in social welfare functions (Piacquadio, 2017).

Axiom (4) requires that DM's choice is *scaling invariant* in some contexts, i.e. that (in some contexts) DM's preferences over any two options are robust to re-scaling the outcome vectors associated with these options by the same factor. In dictator games, for example, such rescaling corresponds to increasing the pie size. Formally, we say that choice in context π^0 is scaling invariant if $\pi^0(x) = \lambda\pi^0(x')$ and $\pi^0(y) = \lambda\pi^0(y')$ for some $\lambda > 0$ implies that $\pi^0(x) \succsim_{\pi^0} \pi^0(y) \Leftrightarrow \pi^0(x') \succsim_{\pi^0} \pi^0(y')$. We shall use $\Pi^0 \subseteq \Pi$ to denote the set of contexts in which choice is scaling invariant. Note that Axiom (4) will not require choice to be scaling invariant in all contexts (though it may be). It requires that, for each context π, at least one context with the same "net default" $\pi(0) - \min \pi$ but possibly different background income $\min \pi$ or range of outcomes exhibits scaling-invariant preferences.

Before we discuss scaling invariance, let us state these first four axioms formally. To simplify the notation, let us introduce a similarity relation for contexts. We write, for any two $\pi, \pi' \in \Pi$, that $\pi \sim \pi'$ if the preferences over outcomes in the two contexts are observably equivalent, i.e. if for all x, y, x', y' such that $\pi(x) = \pi'(x')$ and $\pi(y) = \pi'(y')$, we have $\pi(x) \succsim_\pi \pi(y)$ implies $\pi'(x') \succsim_{\pi'} \pi'(y')$, and if there exists at least one such set $\{x, y, x', y'\}$.

Assumption 2 (Axioms I)**.** For all $\pi \in \Pi$ and all $x, y \in X$:
well defined, i.e. for all $\pi, \pi' \in \Pi$, if $\pi' = \pi + c$ then $\pi'(x) = \pi(x) + c$ for all $x \in X$.

(1) *Weak order:* $\succsim_\pi$ is complete and transitive.

(2) *Continuity:* If $\pi(x) \succ_\pi \pi(y)$, there exists $\varepsilon > 0$ such that $\pi(x') \succ_\pi \pi(y')$ for all $x' : d(x',x) < \varepsilon$ and all $y' : d(y',y) < \varepsilon$.

(3) *Separability:* For all $x', y' \in X$ such that $\pi_{-i}(x) = \pi_{-i}(x')$ and $\pi_{-i}(y) = \pi_{-i}(y')$, as well as $\pi_i(x) = \pi_i(y)$ and $\pi_i(x') = \pi_i(y')$, we have $\pi(x) \succsim_\pi \pi(y)$ iff $\pi(x') \succsim_\pi \pi(y')$.

(4) *Weak scaling invariance:* There exists $\pi^0 \in \Pi^0$ such that $\pi^0(0) - \min \pi^0 = \pi(0) - \min \pi$.

The existence of contexts exhibiting scaling-invariant choice is supported by a host of meta analyses showing that scaling differences between experiments are indeed choice-irrelevant overall. This applies to scaling outcomes in dictator games (Carpenter et al., 2005; Engel, 2011), ultimatum games (Oosterbeek et al., 2004; Cooper and Dutcher, 2011), trust games (Johnson and Mislin, 2011), and choice under risk (Wilcox, 2008, 2011, 2015). At the neuro-physiological level, scaling invariance is implied by *adaptive coding* (Padoa-Schioppa and Rustichini, 2014): The best option always has the maximal firing rate and the worst option always has the minimal firing rate, implying that choice is independent of scale after a transition period where the neuronal firing rate adapts to the scale of the decision problem. Adaptive coding enables efficient adaptation to choice environments subject to the physical limitations in neuronal firing rates (Tremblay and Schultz, 1999; Camerer et al., 2017) and implies that the utility function is homothetic, which is satisfied by a broad range of utility functions discussed in the behavioral literature, including CES altruism, inequity aversion, prospect theoretical utilities, and nested CES functions.

Axioms (5)–(8) capture how preferences react to changes of the context information in π. We invoke either Axiom (5) or Axioms (6)–(8). On the one hand, Axiom (5) ("broad bracketing") requires that preferences over outcomes are insensitive to changes of the context, i.e. to changes of range of outcomes, background income, and default. While this axiom does not seem to be compatible with existing evidence, it provides a benchmark for our discussion.

On the other hand, Axioms (6)–(8) describe a non-trivial way in which DM may respond to changes in context. This is the main novelty of our analysis. While generalizations and alternative approaches toward capturing context dependence are easily conceivable given the results reported below, we think that the context-dependence axioms imposed here seem appealing in relation to the existing literature. Briefly, Axiom (6) captures reactions to changes in the background income vector, Axiom (7) captures reactions to changes of the context in other respects (e.g. changes of ranges of outcomes, of outcome scale, or of the default), and Axiom (8) requires continuity of reactions to changes in context.

To begin with, Axiom (6) (*narrow bracketing*) requires that DM factors out changes in background income, i.e. that she focuses on the net effects of the decision task at hand. This relates to existing literature, e.g. Read et al. (1999), where narrow bracketing refers to the phenomenon that concurrent decision problems are treated independently by decision makers, implying that other tasks simply provide a background income that is factored out. There is ample emprical evidence in favor of narrow bracketing (e.g. Read et al., 1999, Rabin and Weizsäcker, 2009, Simonsohn and Gino, 2013) and of the more general observation that behavior tends to be independent of socio-economic background variables such as income or wealth in experiments (Gächter et al., 2004; Bellemare et al., 2008, 2011) and in general (Easterlin, 2001). In addition, the aforementioned evidence for adaptive coding implies narrow bracketing as well. Specifically, Padoa-Schioppa (2009) show that the baseline activity of the cell encoding the value of a given object generally represents the minimum of the value range and that the upper bound of the activity range of this "value cell" represents the upper bound of the value range. Thus, background utility is factored out and choice satisfies narrow bracketing simply as a result of the physical limitations in neuronal firing. We demonstrate in the following that narrow bracketing implies reference dependence of preferences.

The remaining two axioms characterize reactions to changes in context that are not simply changes in background incomes. We do not impose a specific structure, but merely *additive compensability* (Axiom (7)) and *context continuity* (Axiom (8)). Compensability requires that we can

compensate the decision maker when moving from one context to another by providing side payments that adapt the background income vector. We impose that compensability is additive in the sense that compensations are additive when we combine two changes of the context iteratively. We are not aware of experimental tests of compensability. It does, however, relate to the standard notion of equivalent variation and, more generally, to compensatory incomes which have wide and intuitive appeal. Note also that additivity is satisfied for the equivalent variation. Finally, *context continuity* (Axiom (8)) simply extends preference continuity to the domain of contexts. It requires that reactions to small changes in background income and default induce small changes in the preference ordering in the usual sense that a strict preference for one option over another option is not reverted if the change in background income and default is sufficiently small.

Assumption 3 (Axioms II)**.** For all $\pi \in \Pi$ and all $x, y \in X$:

(5) *Broad bracketing (context independence):* $\pi \sim \pi'$ for all $\pi' \in \Pi$.

(6) *Narrow bracketing (limited context dependence):* For all $\pi' \in \Pi$, if $\pi(0) - \min \pi = \pi'(0) - \min \pi'$, then for all $x, x', y, y' \in X$ such that $\pi(x) - \min \pi = \pi'(x') - \min \pi'$ and $\pi(y) - \min \pi = \pi'(y') - \min \pi'$ we have $\pi(x) \succsim_\pi \pi(y)$ implies $\pi'(x') \succsim_{\pi'} \pi'(y')$.

(7) *Additive compensability:* For all $\pi' \in \Pi$ there exists $c \in \mathbb{R}^n$ such that $\pi' \sim \pi + c$. Further, if $\pi' \sim \pi + c'$ and $\pi'' \sim \pi + c''$, then $\pi' + \pi'' \sim 2\pi + c' + c''$.

(8) *Context continuity:* If $\pi(x) \succ_\pi \pi(y)$, there exists $\varepsilon > 0$ such that $\pi'(x') \succ_{\pi'} \pi'(y')$ for all $\pi' \in \Pi$ such that $d(\pi(0) - \min \pi, \pi'(0) - \min \pi') < \varepsilon$ and all (x', y') such that $d(\pi(x), \pi'(x')) < \varepsilon$ and $d(\pi(y), \pi'(y')) < \varepsilon$.

Which utility functions represent preferences of a DM behaving in line with axioms (1)–(4) and either (5) or (6)–(8)? As usual, we say that a preference relation $\succsim_\pi$ is represented by a utility function $u_\pi : X \to \mathbb{R}$ if $\pi(x) \succsim_\pi \pi(y) \Leftrightarrow u_\pi(x) \geq u_\pi(y)$ for all $x, y \in X$. Proposition 1 establishes that, in conjunction with the other axioms, preferences compatible with broad bracketing (Axiom (5)) are represented by CES altruism and preferences compatible with narrow bracketing (Axioms (6)–(8)) are represented by generalized prospect theoretical preferences where reference points are linear functions of background income $\min \pi$ and net default $\pi(0) - \min \pi$.

Definition 2 (CRRA)**.** A value function $v_i : \mathbb{R} \to \mathbb{R}$ is called (β, δ_i)–CRRA if for all $p \in \mathbb{R}$

$$v_i(p) \underset{\beta \neq 0}{=} \begin{cases} p^\beta/\beta, & \text{if } p \geq 0, \\ -\delta_i \cdot (-p)^\beta/\beta, & \text{if } p < 0, \end{cases} \qquad \text{and} \qquad v_i(p) \underset{\beta = 0}{=} \log(p).$$

Definition 3 (Prospect theoretical)**.** A value function $v_i : \mathbb{R} \to \mathbb{R}$ is called $(\beta, \delta_i, w_i, \pi)$–prospect theoretical if, with the (β, δ_i)-CRRA function $\tilde{v}_i$,

$$v_i(p) = \tilde{v}_i(p - r_i(\pi)) \qquad \text{where} \qquad r_i(\pi) = \min \pi_i + \sum_{j \leq n} w_{i,j} \cdot (\pi_j(0) - \min \pi_j).$$

Proposition 1. *Given Assumption 1, there exist* $\alpha \in \mathbb{R}^n$, $\beta \in \mathbb{R}$, $\delta \in \mathbb{R}^n$ *and* $\mathbf{w} \in \mathbb{R}^{n \times n}$ *such that for all contexts* $\pi \in \Pi$, $\succsim_\pi$ *is represented by*

(a) Axioms (1)–(4),(5) $\iff$ $u_\pi(x') = \sum_{i \leq n} \alpha_i \cdot v_i\big[\pi_i(x')\big]$ *where* v_i *is* (β, δ_i)*–CRRA for all* $i \leq n$*,*

(b) Axioms (1)–(4),(6)–(8) $\iff$ $u_\pi(x') = \sum_{i \leq n} \alpha_i \cdot v_i\big[\pi_i(x')\big]$ *where* v_i *is* $(\beta, \delta_i, w_i, \pi)$*–prospect theoretical for all* $i \leq n$*.*

That is, given Axioms 1–4, a DM obeying broad bracketing has context-independent CES preferences (Andreoni and Miller, 2002), reminiscent of multi-dimensional CRRA value functions inflected at 0 (to cover negative outcome values). In contrast, a DM obeying narrow bracketing and compensability has context-dependent preferences (note that the reference points depend on context π) reminiscent of multi-dimensional prospect theoretical value functions, with the reference

point being a simple (linear) function of background income and default. Quadratic formulations of such utility functions have been called fairness-based preferences in the literature (Konow, 2000; Cappelen et al., 2007), a name we therefore adopt here. Note, however, that quadratic utilities as such are not scaling invariant in our sense. Still, the weights w in the reference point function relate to fairness ideals discussed in the literature (Cappelen et al., 2007), as illustrated below.

To provide some intuition for the result, let us outline the central arguments made in the proof. The existence of a continuous utility function that represents $\succsim_\pi$ is implied (for any π) by Axioms 1 and 2. Given this, Separability (Axiom (3)) ensures that an additively separable utility representation exists (Wakker, 1989), i.e. that for any π, the preference ordering $\succsim_\pi$ can be represented by a utility function $u_\pi : X \to \mathbb{R}$ of the form

$$u_\pi(x) = \sum_{i \leq n} v_{\pi,i}\big(\pi_i(x)\big) \tag{1}$$

with continuous value functions $\{v_{\pi,i} : \mathbb{R} \to \mathbb{R}\}_{i \leq n}$. The remaining axioms clarify the functional form of v_π and how v_π depends on π—or more specifically, they induce the S-shaped pattern known from Prospect theory and establish how the reference points depend on the context information included in π (background income, outcome range and default).

By *scaling invariance* (Axiom (4)), we know that for any scaling-invariant context $\pi^0 \in \Pi^0$,

$$u_{\pi^0}(x) = \sum_{i \leq n} v_i\big(\pi_i^0(x)\big) \qquad \text{and} \qquad u_{\lambda\pi^0}(x) = \sum_{i \leq n} v_i\big(\lambda\pi_i^0(x)\big)$$

with $\lambda \in (0,1)$ both represent $\succsim_{\pi^0}$, and both being additively separable, this implies that they are positive affine transformations of one another. Hence, for all $i \leq n$,

$$v_i\big(\lambda\pi_i^0(x)\big) = a_i(\lambda) + b(\lambda) \cdot v_i\big(\pi_i^0(x)\big)$$

for some functions $a_i : \mathbb{R} \to \mathbb{R}$ and $b : \mathbb{R} \to \mathbb{R}_+$. By Assumption 1.3, the value function v_i is defined on an interval of positive length, by Axiom (2) it is continuous, and by 1.4 it is not equal to the constant function, which jointly implies that the unique solutions of this Pexider functional equation (Aczél, 1966) are the power and logarithmic functions defined in Proposition 1.[7]

As a result of narrow bracketing, there exist n-dimensional reference points for each context, captured by a function $r : \Pi \to \mathbb{R}^n$, such that

$$u_\pi(x) = \sum_{i \leq n} v_i\big(\pi_i(x) - r_i(\pi)\big) \tag{2}$$

represent $\succsim_\pi$ for all $\pi \in \Pi$. Further, narrow bracketing implies that the reference points $r(\pi)$ are a function of background income $\min \pi$ and net default $\pi(0) - \min \pi$. Compensability implies that the reference points scale linearly, $r(\lambda\pi) = \lambda\, r(\pi)$, i.e. scaling invariance between contexts, and its additivity implies that reference points satisfy $r(\pi + \pi') = r(\pi) + r(\pi')$. This equation characterizes the reference point function and thanks to context continuity (Axiom (8)) and default richness, its general solution implies that the reference point is a linear function of the net default $\pi(0) - \min \pi$ and background income $\min \pi$, as stated in the proposition.

Finally, a few technical points appear worth noting. The additive representations are unique up to affine transformation, which implies that the weights (α_i) are unique up to scaling. A standard restriction here is to require that (α_i) add up to 1. With narrow bracketing, the utility function

[7] For illustrative purposes, assume v_i is also differentiable and let $a_i = 0$ (which removes the logarithmic solution). That is, $v_i(\lambda\, \pi_i) = b(\lambda) \cdot v_i(\pi_i)$, and after taking logarithms on both sides, we obtain for $\tilde{v}_i = \log v_i$ and $\tilde{b} = \log b$,

$$\tilde{v}_i(\lambda\, \pi_i) = \tilde{b}(\lambda) + \tilde{v}_i(\pi_i) \qquad \Rightarrow \qquad \tilde{v}_i'(\lambda\, \pi_i) \cdot \pi_i = \tilde{b}'(\lambda) \qquad \Rightarrow \qquad \tilde{v}_i'(\pi_i) = \beta/\pi_i$$

after taking the derivative with respect to λ and letting $\lambda = 1$. This differential equation has the solution $\tilde{v}_i(\pi_i) = \beta \log \pi_i + \alpha_i$ and reverting the logarithm we obtain $v_i(\pi_i) = \alpha_i \cdot \pi_i^\beta$.

is equivalently expressed as

$$u_\pi(x) = \sum_{i \leq n} \alpha_i \cdot \big[r_i(\pi(0) - \min \pi) + v_i\big(\pi_i(x) - r_i(\pi(0) - \min \pi)\big)\big], \tag{3}$$

simply adding the reference points in all dimensions (or any other constant; given separability, the utility function is unique up to positive affine transformation). Formulation (3) may appear more intuitive if the reference points differ from zero.[8] If there exists $x \in X$ such that $\pi^0(x) = 0$, where π^0 denotes a scaling-invariant context, then $\beta > 0$ obtains by continuity. Further, if we assume monotonicity, the parameters (α_i, δ_i) are guaranteed to be non-negative. While this appears plausible in many cases, it would rule out some phenomena resembling inequity aversion, the defining characteristic of which is that preferences are non-monotonic in the opponents' outcomes.[9]

3.3 Discussion

To summarize, broad bracketing induces a context-independent reference point of zero, yielding the well-known CES model of altruism in which outcomes are evaluated in absolute terms. This model represents altruism as a concern for the well-being of others if all agent's individual well-beings are represented by CRRA utilities. In contrast, narrow bracketing implies that outcomes are evaluated in relation to reference points $r_i(\pi)$, such that altruism represents a concern for the well-being of others as it is defined in prospect theory. Switching from broad bracketing to narrow bracketing, which has wide empirical support, is thus linked to switching from CRRA utilities to prospect theoretical utilities in the representation of preferences over monetary income, which has even wider empirical support.

While this link and our attempt to derive social preference representations from general behavioral assumptions are novel, our analysis is related to studies of preferences in choice under risk (e.g., Wakker and Tversky, 1993; Skiadas, 2013). Axioms in this branch of literature are similar to Axioms (1)–(4) above, suggesting the possibility of constructing a general, unified foundation of behavior. In particular, analyses of choice under risk generally work with existence of a weak order, continuity, and an independence assumption yielding additive separability across possible outcomes. Skiadas (2016) shows that a system of axioms including scaling invariance implies a form of CES preferences that is similar to CES altruism characterized based on Axioms (1)–(5), while one including translation invariance implies exponential rather than power utilities resembling constant absolute risk aversion. His results suggest that scaling invariance and translation invariance are mutually exclusive in axiomatic foundations, although both tend to be confirmed in behavioral meta studies. This conflict is resolved with our context-based approach which imposes translation invariance between contexts. The distinction of contexts and the "context dependence" Axioms (6)–(8) including narrow bracketing are a key difference of our analysis compared to previous work. A difference that turns out to be substantial as it allows us to substitute narrow bracketing for translation invariance, which endogenously yields reference dependence while maintaining scaling invariance, and thus obtain social preference representations compatible with previous work on choice under risk and the wide range of observations on distributive decisions discussed next.

Narrow bracketing implies reference dependence and in conjunction with additive compensability it yields the testable predictions that reference points are linear functions of background income and default. This result generalizes existing axiomatic foundations of prospect theoretical utilities, which explicitly assume existence of a reference point, where the reference point is either an exogenously defined payoff vector (Wakker and Tversky, 1993; Wakker and Zank, 2002) or a well-defined option (Schmidt, 2003). Further, we link narrow bracketing and reference

[8] It expresses the idea that meeting one's reference point implies a utility exactly equal to the reference point (in case the value function is the power function in Proposition 1). Thus, for example, an individual being \$10 short of their reference point \$1,000,000 would enjoy a higher utility than an individual being \$10 short of their reference point \$20.

[9] For example, inequity averse subjects prefer $(10,9)$ over $(11,20)$, or $(0,0)$ over $(1,9)$. Without monotonicity, fairness-based altruism can capture such preferences, and in this way, it can also capture rejections in ultimatum games.

dependence based on axioms not related specifically to altruism or giving, underlining the link's generality and corroborating the observation that both narrow bracketing and reference dependence build on a wealth of empirical evidence (outside prospect theory, see for example Kőszegi and Rabin, 2007, 2009, for discussion). Reference points as characterized above are functions of the default (Kahneman et al., 1991) as central benchmark $\pi(0)$, and they may equate with social norms or expectations (Kőszegi and Rabin, 2006) anticipated by DM, as discussed next.

4 Implications for giving, taking and sorting

In this section, we characterize the distributive decisions made by fairness-based altruists and analyze how they relate to the observations made in experiments. Adopting the notation of distribution games (Definition 1), fairness-based altruism as characterized in Proposition 1 is represented in game $\Gamma = \langle B_1, B_2, P_1, t\rangle$ as

$$u_\Gamma(p_1) = \frac{1}{\beta} \times \begin{Bmatrix} (p_1 - r_1)^\beta & \text{if } p_1 \geq r_1 \\ -\delta(r_1 - p_1)^\beta & \text{if } p_1 < r_1 \end{Bmatrix} + \frac{\alpha}{\beta} \times \begin{Bmatrix} (t(B - p_1) - r_2)^\beta & \text{if } p_2(p_1) \geq r_2 \\ -\delta(r_2 - t(B - p_1))^\beta & \text{if } p_2(p_1) < r_2 \end{Bmatrix}.$$

As above, α represents the degree of altruism, δ is the degree of loss-aversion, and β captures the trade-off between efficiency and equity concerns; $\frac{1}{1+\beta}$ is the elasticity of substitution between dictator's and recipient's well-being. Without loss of generality, we assume that δ is the same for both players, and for notational simplicity, we skip the limiting case $\beta = 0$ here. With the convention that the benchmark option $\pi(0)$ is the status quo B, the reference points are

$$r_1 = \min p_1 + w_1 \cdot (B_1 - \min p_1) + w_2 \cdot (B_2 - \min p_2/t),$$
$$r_2 = \min p_2 + w_2 \cdot t \cdot (B_1 - \min p_1) + w_1 \cdot (t \cdot B_2 - \min p_2),$$

again assuming symmetry for notational convenience. Intuitively, each player's reference point is her minimal payoff $\min p_i$ ("background income") plus share $w_1 \in [0,1]$ of the amount she contributes (by default) to the cake to be redistributed ($B_i - \min p_i$) and share $w_2 \in [0,1]$ of the amount her partner contributes to the cake ($B_j - \min p_j$).[10] We assume $w_1, w_2 \geq 0$, i.e. that each player expects to be allocated weakly more in absolute terms as the cake increases (whoever contributed to this increase), and $w_1 \geq w_2$, i.e. that each player believes to be weakly more entitled to get some share of her own contribution than of her partner's contribution.

Assumption 4. $w_1 \geq w_2 \geq 0$.

This model contains status-quo-based reference points ($w_1 = w_2 = 0$) and expectations-based reference points ($w_1 + w_2 = 1$) as notable special cases, and with $w_1 + w_2 \in (0,1)$ all convex combinations thereof. More generally, following Cappelen et al. (2007), the weight w_1 on the own default payoff indicates to which extent DM agrees with the libertarian fairness ideal, demanding no redistribution in relation to the default ($w_1 = 1$, $w_2 = 0$), and the weight on the other's default payoff w_2 indicates the degree to which DM agrees with the egalitarian fairness ideal demanding full redistribution eliminating default payoff differences ($w_1 = w_2 = 0.5$). In this sense, w_2 also represents the degree to which DM feels social pressure to redistribute.

Dictators are fairness-based altruists denoted as $\Delta = (\alpha, \beta, \delta, w_1, w_2)$. Our theoretical analysis will exploit several "regularity" assumptions to ensure that preferences are well-behaved. We assume that dictators are imperfectly altruistic ($0 \leq \alpha \leq 1$), imperfectly efficiency concerned ($0 < \beta < 1$), and weakly loss averse ($\delta \geq 1$). Both $0 < \beta < 1$ and $\delta \geq 1$ are standard assumptions in, for example, prospect theoretical analyses, ensuring S-shaped utilities and avoiding loss seeking, which we therefore adopt as well. Weak altruism ($\alpha \leq 1$) is a standard assumption in analyses of social preferences and $\alpha \geq 0$ is assumed without loss of generality as egoism ($\alpha = 0$) is equivalent

[10] While these reference points follow from our axiomatic characterization, previous work such as Cox et al. (2016) assume similar reference points in distributive decisions.

to spite ($\alpha < 0$) in the games we analyze. Further, we assume that reference points are satisfiable ($w_1 + w_2 \leq 1$) in the sense that an option exists where both players are allocated an amount that is at least equal to their reference point, to reduce the number of cases with corner solutions, which further simplifies the exposition.

Definition 4. Dictator $\Delta = (\alpha, \beta, \delta, w_1, w_2)$ is called **regular** if she exhibits imperfect altruism ($0 \leq \alpha \leq 1$), weak efficiency concerns ($0 < \beta < 1$), loss aversion ($\delta \geq 1$), and satisfiability ($w_1 + w_2 \leq 1$).

Proposition 2 formally characterizes giving of fairness-based altruists to provide the basic intuition. Our subsequent result will explore the relations to the stylized facts discussed above.

Proposition 2. *In a given distribution game Γ almost all regular dictators Δ can be classified as follows. Dictators with $\alpha^{1/\beta} < 1/t$ choose*

$$(p_1^*, p_2^*) = \begin{cases} \left(\frac{tB + c_\alpha r_1 - r_2/t}{c_\alpha + 1}, \frac{t c_\alpha (B - r_1) + r_2}{c_\alpha + 1}\right), & \text{if } \delta > \delta^+ \quad \textbf{(interior solution)} \\ (\max p_1, \min p_2), & \text{if } \delta < \delta^+ \quad \textbf{(egoistic solution)} \end{cases}$$

while dictators with $\alpha^{1/\beta} > 1/t$ choose

$$(p_1^*, p_2^*) = \begin{cases} \left(\frac{tB + c_\alpha r_1 - r_2/t}{c_\alpha + 1}, \frac{t c_\alpha (B - r_1) + r_2}{c_\alpha + 1}\right), & \text{if } \delta > \delta^- \quad \textbf{(interior solution)} \\ (\min p_1, \max p_2), & \text{if } \delta < \delta^- \quad \textbf{(altruistic solution)} \end{cases}$$

with

$$\delta^+ := c_\alpha^{\beta-1} \left(\left(\frac{t(B - r_1)}{r_2} \right)^\beta - (c_\alpha + 1)^{1-\beta} \left(\frac{t(B - r_1)}{r_2} - 1 \right)^\beta \right)$$

$$\delta^- := c_\alpha^{1-\beta} \left(\frac{tB - r_2}{t r_1} \right)^\beta - (c_\alpha + 1)^{1-\beta} \left(\frac{tB - r_2}{t r_1} - 1 \right)^\beta$$

and $c_\alpha := (\alpha t^\beta)^{\frac{1}{1-\beta}}$.

That is, there are up to three types of fairness-based altruists: some give nothing or take all (choosing the lower bound), some give a bit (choosing an interior solution), and some give all (choosing the upper bound). In the interior solution, both reference points are satisfied, which implies that many possible decisions can be ruled out. Further, the types of fairness-based altruists are defined using simple thresholds (δ^-, δ^+) in terms of the degree of loss aversion δ. This allows us to rank dictators by their propensity to choose either of the corner solutions. While dictators with a low degree of loss aversion δ tend to have a high propensity to choose a corner solution, evaluating the extra costs of not satisfying a reference point as low, dictators with a high degree of loss aversion tend to pick an interior solution. The type of corner solution chosen by dictators with low δ depends on their degree of altruism α, the welfare function curvature β, and the transfer efficiency t. The altruistic corner solution becomes relevant only in games with efficiency gains from giving ($t > 1$) for dictators who have relatively high altruism weights α and/or strong efficiency concerns (high β). The thresholds (δ^+, δ^-) for actually choosing either corner solution when it is relevant have intuitive comparative statics in the preference parameters. The higher the altruism weight α, the lower the maximum δ for which the egoistic corner solution is chosen and the higher the maximum δ for which the altruistic corner solution is chosen. The stronger the dictator's efficiency concerns (the higher β), the higher the maximum δ for which an efficient corner solution is chosen and the lower the maximum δ for which an inefficient corner solution is chosen.

Since the interior solution and the thresholds δ^+ and δ^- are continuous in the game parameters $\langle B_1, B_2, P_1, t \rangle$ we can also characterize the comparative statics of behavior across different distribution games. The interior solution has very intuitive comparative statics in this respect: The recipient's payoff is decreasing in the dictator's reference point r_1, increasing in the recipient's reference point r_2 and budget B, and increasing in the transfer rate t. In conjunction with the similarly

intuitive comparative statics of the thresholds δ^+ and δ^-, this directly predicts the stylized facts observed in the literature (Table 1). Proposition 3 establishes this formally, a detailed discussion follows. As above, we say a dictator is a "giver" if she transfers some of her endowment to the recipient, she is a "taker" if the net-transfer is negative, and comparing two games, we say that the range of taking options is extended if $B = B_1 + B_2$ is held constant but the maximal dictator transfer $\max p_1$ increases.

Proposition 3. *Assume dictators $\Delta = (\alpha, \beta, \delta, w_1, w_2)$ are randomly distributed in $\mathbb{R}^5$ such that dictator Δ has positive density if and only if dictator Δ is regular. All "stylized facts" are implied.*

1. ***Non-convexity*** *In all games with $P_1 = [0, B]$, some dictators have non-convex preferences.*
2. ***Taking options reduce giving both at the extensive and intensive margin*** *Introducing a taking option turns some initial givers into takers and reduces average amounts given.*
3. ***Incomplete crowding out*** *Reallocating initial endowment from the dictator to the recipient results (in expectation) in a payoff increase for the recipient.*
4. ***Efficiency concerns*** *The recipient's payoff is weakly increasing in the transfer rate.*
5. ***Reluctant sharers*** *When an outside option is introduced, some initial givers switch to that option while the behavior of dictators who sort into the game stays unaffected.*
6. ***Social pressure givers*** *Ceteris paribus, higher susceptibility to social pressure (higher w_2) implies higher transfers in the interior solution but also a higher propensity to choose the outside option in a sorting game.*

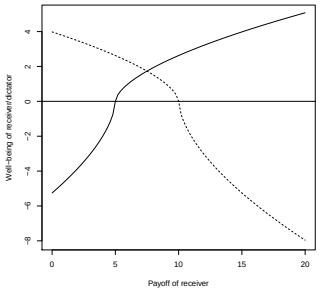
Givers who become takers: Non-convexity of preferences One of the most distinctive characteristics of fairness-based altruism is the implied non-convexity of preferences. This non-convexity has important consequences for our model's theoretical predictions across distribution games that differ in the dictator's choice set, in particular comparing games with generalized endowments to taking games. The nature of non-convexity and its consequences are illustrated in Figure 1.

We consider a distribution game in which the dictator is asked to allocate a budget of 20 tokens between herself and the recipient at a transfer rate of $t = 1$. Suppose that the reference points are $r_2 = 5$ for the recipient and $r_1 = 10$ for the dictator. Figure 1a depicts the trade-off that the dictator faces between her own and the recipient's welfare. The more the dictator allocates to the recipient, the higher is the recipient's welfare (solid curve) but the lower is the dictator's own welfare (dashed curve). The individual welfares are steeper the closer the players are to their respective reference points. For recipient payoffs between 5 and 10, both the recipient and the dictator are in the gain domain, i.e. they achieve payoffs at least as high as their respective reference points, whereas for all other allocations one of them is in the loss domain. Figure 1b depicts the dictator's utility if her weight on the recipient's welfare is $\alpha = 0.3$. This dictator's utility function reaches its maximum at the interior solution where the transfer slightly exceeds the recipient's reference point. Figure 1c depicts the utility of a slightly less altruistic dictator ($\alpha = 0.2$). This dictator's optimal choice is the egoistic (corner) solution of allocating nothing to the recipient.

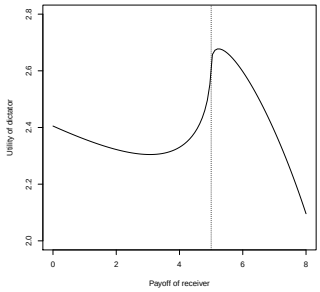
The S-shaped form of the individual welfare function implies that the deeper the recipient moves into the loss domain, the lower the marginal reduction in recipient welfare for any further token not allocated to him. In conjunction with weak altruism and the correspondingly S-shaped dictator welfare, this implies that dictator utility is not quasi-concave—it bends upwards once the recipient is sufficiently far below his reference point. Ceteris paribus, the lower the weight α that the dictator assigns to the recipient's welfare, the earlier this minimum is reached and the more likely it is that the dictator's utility from choosing the lower bound exceeds her utility in the interior solution. As a result, dictator behavior is not generally continuous in the game parameters, which predicts the "preference reversals" observed in taking games.

Figure 1: Non-convexity of preferences and implications in taking games

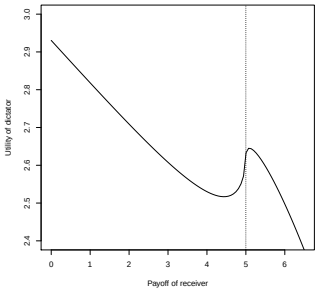
(a) Welfares of recipient (solid) and dictator (dashed) with $\beta = 0.6$ and $\delta = 2$



(b) Utility of a dictator with $\alpha = 0.3$



(c) Utility of a dictator with $\alpha = 0.2$



Note: The dictator can choose to allocate x tokens to the recipient, where $x \in [0, 20]$. The transfer rate is $t = 1$. The recipient's reference point is $r_2 = 5$ while the dictator's reference point is $r_1 = 10$. The dashed lines in (b) and (c) mark the recipient's reference point.

To see this, have another look at Figure 1c, now assuming the recipient's reference point equates with his endowment ($B_2 = 5$ and $B_1 = 15$). That is, if the dictator allocates, say, 4 to the recipient, then she actually takes from his endowment. For simplicity, also assume that reference points are invariant to changes in the dictator's choice set (reference point movements are covered in the subsequent discussion) and suppose the dictator cannot take from the recipient's endowment. In this case, the dictator cannot implement an allocation with a recipient payoff below his reference point, to the left of the vertical dotted line in Figure 1c, and chooses the interior solution to the right. Now, as we extend the option set by allowing for taking one token from the recipient, allocations to the left of the vertical dotted line become admissible. Initially, upon extending the option set, the dictator's utility at the lower bound is decreasing. The recipient's welfare drops sharply and the dictator is concerned for his welfare. Upon further extending the option set into the taking domain, the dictator's utility reaches a minimum and starts to increase again. Eventually, the dictator prefers the lower bound to the interior solution and jumps to taking as much as possible. Such a "preference reversal" cannot be observed for the more altruistic dictator in Figure 1b as long as the recipient's payoff is restricted to be non-negative.

Decreasing the lower bound decreases expectations: Taking options Introducing a taking option decreases the recipient's minimal payoff, i.e. his background income. Regardless of whether the recipient has status-quo-based or expectations-based reference points, or a convex combination from the general class in Assumption 4, the recipient's reference point will consequentially decline. The reduction in the recipient's minimal payoff at the same time raises the surplus $B_2 - \min p_2/t$ he contributes, but generically (for all $w_1 < 1$) the first effect dominates. Loosely speaking, the recipient will be happy with less. In turn, the dictator's reference point weakly increases through her partial claim to the increasing surplus contributed by the recipient (if $w_2 > 0$). That is, after introducing taking options, the dictator asks for more. Both effects directly imply, at the intensive margin, that the dictator transfers less in the interior solution, which has the obvious comparative statics in reference points by Proposition 2. In addition, as the lower bound declines, defecting towards the lower bound becomes more attractive for the dictator (recall Figure 1) and with the increase of the own reference point, the interior solution becomes less attractive. As a result, at the extensive margin, dictators are more likely to pick the lower bound, and across the population, the share of regular dictators who choose the lower bound increases while the share of regular dictators who choose the interior solution decreases.

Shifting surplus claims: Generalized endowments Assume part of the dictator's endowment is reallocated to the recipient and the dictator cannot take any of it back, i.e. her budget correspondingly declines. Then, the dictator's background income is constant but the surplus she contributes ($B_1 - \min p_1$) decreases, while the recipient's background income increases and his surplus remains constant. As a result, the dictator's reference point declines and the recipient's reference point increases. By the comparative statics of the interior solution, the dictator thus allocates less to herself and more to the recipient at the interior solution, implying incomplete crowding out of endowment reallocations.

Avoiding high recipient expectations: Sorting options Lazear et al. (2012) call a dictator a "reluctant sharer" if she transfers a positive amount in a standard dictator game but sorts out when possible. That is, her utility from the interior solution is lower than her utility from the outside option $(\tilde{p}_1, 0)$—assuming the recipient is not informed about the dictator and her options if she sorts out. Remaining uninformed if the dictator sorts out, from the recipient's perspective literally nothing happens, both reference point and payoff are zero, and he remains welfare neutral. This removes the negative externality imposed by the recipient's expectations and may therefore be preferable for the dictator. To see this, assume reference points are just "satisfiable", i.e. $B = r_1 + r_2/t$, and the dictator chooses to satisfy them in the standard dictator game (as opposed to choosing the lower bound). The interior solution generates zero surplus for either player and consequentially zero utility. Then, sorting out is strictly preferable whenever $\tilde{p}_1 > r_1$. If we set $\tilde{p}_1 = B_1$ and start declining it, as in the experiment of Lazear et al. (2012), the condition $\tilde{p}_1 > r_1$ is first violated for dictators with high reference points r_1, who transfer the least at the interior solution. These players are thus predicted to sort in first, regardless of how subjects mix status quo and expectations forming reference points, which corroborates the observation of Lazear et al. (2012) that the least generous dictators sort back in first.

5 Implications for giving: Quantitative assessment

In this section, we quantitatively test fairness-based altruism on actual data from the experiments discussed above. We examine whether fairness-based altruism indeed helps improve our understanding of giving in a statistically significant manner. Besides evaluating significance, this allows us to address three potential concerns: Is the gain larger than two additional degrees of freedom (the reference points) allow to achieve anyways? Does it matter whether these degrees of freedom are spent on defining reference points, as predicted above, or perhaps on warm glow and cold prickle, or envy and guilt, as suggested in the literature? Do the additional degrees of freedom facilitate overfitting?

Arguably, the match of theoretical predictions and empirical stylized facts for all distributions of reference points given "regularity" of dictator preferences strongly suggests that fairness-based altruism does capture giving reliably without the necessity of fine-tuning parameters. Yet, additional degrees of freedom tend to be an obstacle to robust fit (Hey et al., 2010). To address this potential obstacle directly, our analysis will emphasize predictive adequacy over descriptive adequacy. For the lack of comparable analyses in the existing literature, we include a number of well-known models as benchmarks to provide context.

5.1 The data

Table 2 summarizes the experiments we re-analyze. All of them are seminal studies run for the purpose of characterizing preferences underlying giving, rendering them adequate also for our purpose of validating utility representations of preferences. In total, we analyze behavior across 9 experiments, 83 treatments, and 6500 observations. In relation to comparable studies of model validity, e.g. on choice under risk, this represents a very comprehensive data set, promising reliable results.

Table 2: The experiments re-analyzed to verify model adequacy

		Abbreviation	#Treatments	# Subjects	#Observations
Dictator games	Andreoni and Miller (2002)	AM02	8	176	1408
	Harrison and Johnson (2006)	HJ06	10	56	560
Generalized endowments					
	Cappelen et al. (2007)	CHST07	11	96	190
	Korenok et al. (2012)	KMR12	8	34	272
	Korenok et al. (2013)	KMR13	18	119	2142
Taking (and generalized endowments)					
	List (2007)	List07	3	120	120
	Bardsley (2008)	Bard08	6	180	180
	Korenok et al. (2014)	KMR14	9	106	954
Sorting	Lazear et al. (2012)	LMW12	8	94	518
Aggregate		Pooled	83	981	6578

To our knowledge, our data set includes all experiments on distribution games as analyzed above, i.e. with generalized endowments, taking, or sorting options, complete information, at least three treatments, manual entry of choices, and freely available data sets. The focus on experiments with at least three treatments facilitates statistically informative likelihood ratios but it precludes small experiments, most notably a seminal paper on sorting (Dana et al., 2006). The focus on games with complete information facilitates a unified theoretical treatment but precludes field experiments on charitable giving (such as DellaVigna et al., 2012) and experiments on moral wiggle room (Dana et al., 2007; van der Weele et al., 2014). The focus on games with manual choice entry simplifies out-of-sample predictions but precludes experiments with graphical user interfaces (Fisman et al., 2007). Finally, the focus on games with freely available data sets precludes the inclusion of experiments with real-effort tasks preceding a dictator game. However, as reviewed above, the main patterns in real-effort games resemble those in distribution games with generalized endowments and windfall budgets, three of which are included.

A notable difference between the analyzed distribution game experiments concerns the language used in the instructions for assigning the players' endowments. In standard dictator games (e.g. AM02 and HJ06), direct assignments are avoided by stating that "a number of tokens is to be divided", while in taking games (e.g. List07, Bard08, and KMR14), endowments are explicitly assigned prior to the choice task. This may provoke status quo and endowment effects (Samuelson and Zeckhauser, 1988; Kahneman et al., 1991) but to our knowledge it has not been discussed as a behavioral confound in preference analyses of (generalized) dictator games. Table 4 in the appendix reviews the relevant passages in the experimental instructions and distinguishes between neutral language, where specific assignments of the endowments to either of the players are avoided, and loaded language, where initial endowments are specifically assigned or otherwise attributed to either of the players. Neutral language is typically used in standard dictator game experiments (AM02 and HJ06) and in sorting games (LMW12). Loaded language is typically used in experiments with generalized endowments or taking options. The hypothesis that such language differences affect the distribution of reference points and thus induce endowment effects as observed in other studies will be verified below and will be taken into account throughout the entire analysis.

5.2 Heterogeneity and consistency of reference points

For the following analysis, we use the simplest formulation of reference points that seems conceivable, simplifying even in relation to Assumption 4, in order to rule out any biases in the results

due to choosing functional forms.

Definition 5 (Simplified reference points)**.** Given game $\Gamma = \langle B_1, B_2, P_1, t \rangle$, the two players' reference points are

$$r_1 = w_1 \cdot B_1 + w_2 \cdot tB_2, \qquad r_2 = w_1 \cdot tB_2 + w_2 \cdot B_1.$$

Since the qualitative results hold regardless of the distribution of reference points, the specific assumption used here is largely irrelevant, but for any test of the model, some specification has to be used. The robustness checks in Appendix C explicitly show that alternative functional forms mapping endowments to reference points yield results very similar to those reported here. As above, we assume that they satisfy $w_1 \geq w_2$ such that subjects put higher weight on the role they end up playing in case their decision turns out to be payoff relevant, and we continue to allow that the weights $w_1, w_2 \in [0,1]$ do not necessarily add up to 1. The latter allows that subjects may be both altruistic givers and social pressure givers, thereby capturing the types observed by DellaVigna et al. (2012). Specifically, we speak of altruistic givers if $w_1 + w_2 < 1$, in which case satisfiability of reference points is fulfilled and dictators tend to pick (if $\alpha > 0$) interior solutions giving more than necessary to fulfill the recipient's reference point. In contrast, we speak of social-pressure givers if $w_1 + w_2 \geq 1$, which obtains if w_2 is sufficiently large, as the dictator is then unable to give more than "necessary" to both players, implying that she gives only to satisfy the reference points as good as possible. We fix the loss-aversion parameter at the conventional value $\delta = 2$ to remove a degree of freedom.

First, we examine heterogeneity of reference points within experiments (i.e. within subject pools) and consistency of reference point distributions across experiments (i.e. types of dictator games). We begin with examining consistency across experiments. For, the differences in the language used when assigning endowments potentially preclude consistency across experiments, which might render the subsequent robustness analysis futile. Further, it would limit applicability of reference dependent concepts such as fairness-based altruism, or indeed any existing concept, to understand the behavioral reasons for differences in giving across experiments.

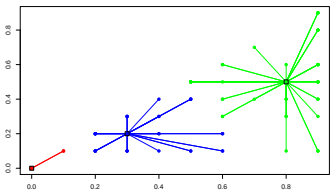
Formally, we estimate the individual reference points of all subjects in the largest experiment from each class of games: dictator games (AM02), games with generalized endowments (KMR13), sorting games (LMW12), and taking games (KMR14). To be precise, we estimate all four individual preference parameters for all subjects, as reference points cannot be estimated without controlling for altruism α and efficiency concerns β, but in the present subsection, we focus on the distributions of reference points. As the estimation procedure is standard maximum likelihood all details on optimization algorithms, generation of starting values, and cross-checking to ensure global optimality of estimates are relegated to the appendix. After estimating the reference point weights (w_1, w_2) for all subjects, we evaluate their structure in a cluster analysis by affinity propagation (Dueck and Frey, 2007). Figure 2 provides the results.

Consistently across data sets, three clusters of subjects are identified. The clusters tend to be of comparable size across experiments, each comprising at least 20% of the subjects in each case. In all cases, there is one group of subjects with endowment-independent reference points ($w_1 \approx w_2 \approx 0$), one group of subjects with "satisfiable" reference points where weights add up to less than one ($w_1 + w_2 < 1$), and one group of subjects with "excessive" reference points where weights add up to more than one ($w_1 + w_2 \geq 1$). The center of the second group moves a little between studies, but overall, the centers and sizes of the clusters are remarkably robust—and they fit received findings in the literature. The first group contains the "egoistic" subjects maximizing their pecuniary payoffs, a group comprising around one third of the subjects in all dictator game experiments. The members of the second and the third group comprise subjects that transfer tokens to the recipients either out of largely altruistic concerns (second group) or out of perceived social pressure (third group)—and further corroborating DellaVigna et al. (2012), these groups are similarly large.[11]

[11] Members of both the second and the third group react to the endowments induced via the experimental design. The difference is that the reference points of members in the second group do not eat up the entire budget, while the reference

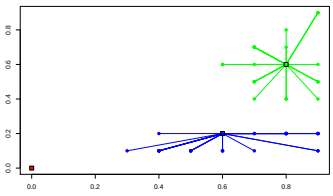
Figure 2: Distribution of reference point weights across types of dictator games

Dictator games (AM02)



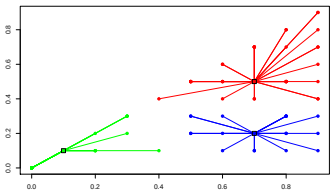
	w_1	w_2	Size	
Cluster 1	0.002	0.002	104/176	59%
Cluster 2	0.359	0.203	32/176	18%
Cluster 3	0.77	0.503	40/176	23%

Generalized endowments (KMR13)



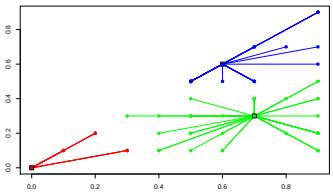
	w_1	w_2	Size	
Cluster 1	0	0	60/119	50%
Cluster 2	0.631	0.144	32/119	27%
Cluster 3	0.774	0.607	27/119	23%

Sorting games (LMW12)



	w_1	w_2	Size	
Cluster 1	0.106	0.084	32/94	34%
Cluster 2	0.71	0.215	20/94	21%
Cluster 3	0.717	0.571	42/94	45%

Taking games (KMR14)



	w_1	w_2	Size	
Cluster 1	0.049	0.032	37/106	35%
Cluster 2	0.69	0.267	39/106	37%
Cluster 3	0.657	0.593	30/106	28%

Note: For the largest experiments from each type of generalized dictator game, all individual reference point weights (w_1, w_2) are estimated, plotted with w_1 on the horizontal axis and w_2 on the vertical axis, and clustered by affinity propagation (Dueck and Frey, 2007). The centers and sizes of the three clusters identified in each case are provided in the respective tables to the right.

Result 1. *Across all four types of dictator games, there are three similarly-sized groups of subjects: subjects with endowment-independent reference points (mostly egoists), subjects with satisfiable reference points ("altruistic givers"), and subjects with non-satisfiable reference points ("social pressure givers").*

5.3 Significance and robustness of fairness-based altruism

Next, if reference dependence is a *robust* behavioral trait, then accounting for it improves both our descriptions and predictions of behavior across contexts. Besides being an informative test statistic, predictive adequacy is important also to improve policy recommendations and guide (behavioral) mechanism design. Given the data sets analyzed here, we can replicate out-of-sample predictions as used in such applications by making predictions across the types of dictator game

points of members in the third group cannot be satisfied jointly. The members of the third group transfer tokens aiming to satisfy both players' reference points as good as possible, and in this sense, they react solely to the social pressure they perceive due to their (subjective) reference points. The members of the second group, however, react significantly weaker to the social pressure (i.e. to induced endowments), thanks to having smaller weights (w_1, w_2) and mainly decide how to transfer the (often substantial) residual amount after satisfying both reference points. In this sense, they are altruistic givers.

experiments listed in Table 2.

In addition, if reference dependence is a *behavioral primitive*, then it improves on alternative ways of providing the implied degrees of freedom. Given the existing literature, there are two arguably natural extensions of CES altruism that have to be considered as benchmark models. The first benchmark extends CES altruism by warm glow and cold prickle, as proposed by Korenok et al. (2014):

$$u(\pi) = (1-\alpha_1-\alpha_2-\alpha_3)\cdot\pi_1^\beta + \alpha_1\ \pi_2^\beta + \alpha_2\cdot|B_1-\pi_1|_+^\beta - \alpha_3\cdot|B_2-\pi_2|_+^\beta, \quad \text{(+ Warm Glow/Cold Prickle)}$$

where $|x|_+$ equates with x if $x>0$ and with 0 otherwise. Thus, $|B_1-\pi_1|_+$ captures the amount transferred by the dictator from her endowment (inducing "warm glow" which is independent of the amount received by the recipient), and $|B_2-\pi_2|_+$ captures the amount taken from the recipient's endowment (inducing "cold prickle"). The other benchmark extends CES altruism by motives of envy and guilt (Fehr and Schmidt, 1999) as proposed by Korenok et al. (2012).

$$u(\pi) = (1-\alpha_1-\alpha_2-\alpha_3)\cdot\pi_1^\beta + \alpha_1\ \pi_2^\beta - \alpha_2\cdot|\pi_1-\pi_2|_+ - \alpha_3\cdot|\pi_2-\pi_1|_+ \quad \text{(+ Inequity Aversion)}$$

An attractive feature of these models is that they also contain four free parameters in total, in this respect equating with fairness-based altruism, which implies that these models can be estimated following the exact same procedure as fairness-based altruism. This way, we can ensure comparability of the results. All the technical details on likelihood maximization and statistical tests are provided in the appendix.

We estimate all models on each of the three largest data sets, i.e. on standard dictator games (AM02), on games with generalized endowments (KMR13), and on games with taking options (KMR14), and predict behavior in all data sets listed in Table 2.[12] The results are summarized in Table 3. For completeness, we also provide the "Descriptive Adequacy", which is the Akaike information criterion of the in-sample fit, i.e. the sum of the absolute value of the log-likelihood and the number of parameters (in-sample, every reference point of every subject counts as a free parameter). Given the large number of parameters, the descriptive adequacy is of limited informational content on its own.

Our focus is on the "Predictive Adequacy", which is reported both on aggregate (column "Predictive Adequacy") and segregated by type of dictator game to be predicted (sets of columns "Details on predictions of ..."). In all cases, descriptive and predictive adequacies are reported for each of the four models discussed so far, payoff-based CES altruism, the extensions additionally allowing for either warm glow and cold prickle or envy and guilt, and the fairness-based altruism model. In addition, we report results from a robustness check allowing for variations in the strength of assignments of endowments, the model "Welfare based (adj)" that we discuss below. Finally, in the lower part of Table 3, all the numbers in the upper part are aggregated across all three in-sample data sets to provide the overall picture.

Descriptive adequacy Briefly, let us look at the in-sample fit (column "Descriptive Adequacy"). On aggregate, all generalized models significantly improve on the payoff-based CES model despite accounting for the additional parameters using AIC. The proposed model of fairness-based altruism is unique in that it improves highly significantly upon CES in all three contexts. In this sense it represents the only robustly fitting model. Yet, the observation that on aggregate all three models do so suggests that they might all capture differently important but significant facets of behavior. If so, this will show in their predictive adequacy.

[12] We do not consider predictions based on estimates from the sorting game experiment of LMW12, as their experimental design varies neither the transfer rate (fixed to 1 : 1) nor the endowments of dictators and receivers, varying only the price for sorting out. This way, the preference parameter β, capturing the preference for efficiency and equity, is not identified and predictions are largely uninformative.

Table 3: Behavioral predictions across types of dictator game experiments

Calibrated on	Altruism is ...	Descriptive Adequacy	Predictive Adequacy	Details on predictions of ...			
				Dictator	Endowments	Taking	Sorting
Dictator Games	Payoff based (CES)	1460.9	8950.5	1343.4	4339	2353.3	914.7
	+ Warm Glow/Cold Prickle	1507.3^{--}	8854.6	1343	4218.4^{+}	2375.2	917.9
	+ Inequity Aversion	1234.6^{++}	8794.8^{++}	1217.1^{+}	4311.7	2360.7	905.3
	Fairness based	1146.6^{++}	8758^{++}	1279.8^{+}	4273.8^{+}	2316.6^{+}	887.7
	Fairness based (adj)	1146.6^{++}	8603.9^{++}	1263.9^{+}	4152.5^{++}	2300.8^{++}	888.2
Gen Endowments	Payoff based (CES)	2896.6	8752.9	4260.4	826.1	2613.8	1052.7
	+ Warm Glow/Cold Prickle	2395.5^{++}	8967.8^{--}	4289.6	954.5^{--}	2649.7	1074
	+ Inequity Aversion	2800.1^{+}	8916.4^{--}	4333.6^{-}	849.9	2663^{--}	1069.9^{--}
	Fairness based	2662.7^{++}	8416.7^{++}	4084.2^{++}	767.9^{+}	2565.9^{+}	998.7^{++}
	Fairness based (adj)	2662.7^{++}	7867.7^{++}	3985.8^{++}	637.1^{++}	2351^{++}	895.4^{++}
Taking Games	Payoff-based (CES)	1482.4	9700.7	3739.3	4466.7	579.7	914.9
	+ Warm Glow/Cold Prickle	1451.8	10252.5^{--}	4263.8^{--}	4408.7	592.8	987.2^{--}
	+ Inequity Aversion	1419.2^{+}	9736.7	3543.3^{++}	4698.2^{--}	576.6	918.5
	Fairness based	1226.4^{++}	9499.7^{+}	3729.2	4343.2	568.5^{+}	858.8^{++}
	Fairness based (adj)	1226.4^{++}	9270.3^{++}	3633^{+}	4232.9^{++}	559.3^{++}	846.6^{++}
Aggregate	Payoff based (CES)	5839.8	27404.1	9343.1	9631.8	5546.8	2882.4
	+ Warm Glow/Cold Prickle	5354.6^{++}	28075^{--}	9896.5^{--}	9581.6	5617.8^{-}	2979.1^{--}
	+ Inequity Aversion	5453.9^{++}	27447.9	9094^{+}	9859.8^{--}	5600.4^{--}	2893.7
	Fairness based	5035.7^{++}	26674.4^{++}	9093.2^{++}	9385^{++}	5451^{++}	2745.2^{++}
	Fairness based (adj)	5035.7^{++}	25740.4^{++}	8883.6^{++}	9023.5^{++}	5212.2^{++}	2631.2^{++}

Note: For each type of dictator game experiment used to estimate the parameters (standard "Dictator games" in AM02, "Generalized endowments" in KMR13, "Taking Games" in KMR14), we report for each of the five models the in-sample fit ("Descriptive Adequacy"), the pooled out-of-sample fit by predicting all other experiments in Table 2 ("Predictive Adequacy"), and the detailed predictive adequacy for each type of experiments as distinguished in Table 2 (the four right-most columns). Plus and Minus signs indicate significance of differences of the Akaike Information Criterion (AIC) for each of the generalizations of the CES model to the CES model. The likelihood-ratio tests (Schennach and Wilhelm, 2016) are robust to misspecification and arbitrary nesting, and we distinguish significance levels of .05 ($^{+}$,$^{-}$) and .01 ($^{++}$,$^{--}$). In all cases, we cluster at the subject level to account for the panel character of the data.

Predictive adequacy Evaluating robustness of the explanatory power (column "Predictive Adequacy") changes the picture substantially. Fairness-based altruism improves on CES' predictions in all contexts, regardless of the data set used for estimation, and mostly significantly so. That is, regardless of the context the model is fitted on and of the class of dictator game experiments to be predicted, the resulting goodness-of-fit is higher than that of the standard CES model, in all 3×4 cases, significantly so in $9/12$ cases, and always on aggregate.[13] The explanatory power of reference dependence in giving may therefore be considered robust.

At the other extreme, extending CES altruism by warm glow and cold prickle predicts behavior better than CES in only $3/12$ cases but worse than CES in $9/12$ cases. On aggregate, the alternative model's predictions are significantly worse than CES, and this obtains although warm glow and cold prickle seem to capture behavior (in-sample) in the case of generalized endowments best. This applies only in-sample, however, even predictions for the other experiments allowing for generalized endowments fit worse than CES (and all other models), suggesting that the extension allowing for warm glow and cold prickle does not capture a robust behavioral trait in the games analyzed here.

Finally, the extension allowing for envy and guilt ("inequity aversion") is in-between with

[13] Note that, as mentioned in the notes to all tables and in the appendix, we use the Schennach-Wilhelm likelihood ratio test throughout (Schennach and Wilhelm, 2016), clustered at the subject level. It is robust to misspecification of models, arbitrary nesting structures, and captures the panel character of the data with multiple observations per subject.

respect to its descriptive and predictive adequacy. While it fits worse than fairness-based altruism in all contexts, both in-sample and out-of-sample, at least it does not overfit on aggregate and thereby it improves on warm glow and cold prickle. That is, on aggregate, accounting for envy and guilt does not yield predictions that are significantly worse than not doing so (as in the standard CES model). Nonetheless, predictions also do not improve on aggregate, suggesting that envy and guilt are actually not robust behavioral traits in giving—they allow to rationalize Leontief choices, but those are not robustly chosen.[14] Corroborating this observation, if we evaluate predictions across all 4×3 cases, inequity aversion's predictions significantly improve on CES in $2/12$ cases, it predicts significantly worse in $4/12$ cases, and overall, its predictive adequacy is slightly worse than the one of the payoff-based CES model.

Result 2. *Fairness-based altruism improves on CES altruism for all types of DG experiments, both descriptively (in-sample) and robustly (out-of-sample) highly significantly. None of the benchmark models does so in more than* $2/12$ *cases, corroborating the theoretical prediction that reference dependence is a causal factor in giving across contexts.*

Table 3 additionally informs on a robustness check accounting for the variation in language used assigning endowments (Table 4 in the appendix). In this robustness check, we allow for homogeneous shifts in weights between experiments, by introducing a free parameter per set of predictions. Assuming the in-sample estimates of the weights are (w_1, w_2), we allow the out-of-sample weights to be (w_1^γ, w_2^γ), where the shift $\gamma \geq 0$ is homogeneous for all subjects. With $\gamma < 1$ all weights increase and with $\gamma > 1$ all weights decrease—reflecting stronger and weaker assignments, respectively. Introducing γ as a free parameter allows us to either strengthen or weaken weights homogeneously for all subjects. Naturally, this has no effect in-sample, but it has substantial effects out-of-sample—amounting to around 1000 points on the log-likelihood scale in total (yielding a drop from 26674.4 to 25740.4). This improvement is highly significant given the low number of additional parameters used, strongly underlining the initial hypothesis that the language used in experimental instructions is highly relevant in shaping behavior. The present analysis is neither suited nor intended to fully clarify the relevance of language used assigning endowments, but changes in language across experiments, which have not been explicitly discussed in the literature on generalized dictator games, are evidently not innocent choices in experimental design. This does not directly affect the above results, since acknowledging language differences as a factor shaping reference points only strengthens the case for fairness-based altruism, but such differences may be acknowledged more explicitly when designing and analyzing future experiments.

6 Conclusion

This paper contributes to the efforts in reorganizing models of the interdependence of preferences (List, 2009; Malmendier et al., 2014) that was initiated by a wave of distribution game experiments generalizing the standard dictator game allowing for non-trivial endowments (Bolton and Katok, 1998; Korenok et al., 2013), taking options (List, 2007; Bardsley, 2008), and sorting options (Dana et al., 2006; Lazear et al., 2012). The new observations were interpreted as being incompatible with observations from standard dictator games and in the existing literature a plethora of approaches have been proposed to capture them: menu dependent preferences and cold prickle to capture taking decisions, warm glow and social norms to capture endowment effects, image concerns and social pressure to capture sorting decisions. Considering this range of proposals simply to organize observations on giving under complete information, robustly applicable models of this most fundamental of economic activities appear to be out of reach (Korenok et al., 2014)—indicating a surprisingly tight bound on economic modeling.

[14] In particular in the games with generalized (non-zero) endowments, the payoff-equalizing "Leontief" option happens to be rarely chosen (Korenok et al., 2013). For example, only $2/116$ subjects in KMR14 are strict Leontief types, whereas around 20% of the subjects are in standard dictator games (see AM02). In this context, predictions assuming that envy and guilt are behavioral factors fit poorly.

We propose an axiomatic approach toward modeling preferences that resolves the persistent puzzles surrounding distributive decisions. It differs from earlier work in four important ways. First, relying on an axiomatic foundation allows us to characterize a general family of utility functions representing interdependence of preferences. This identifies the class of candidate models. Second, we complement the axiomatic analysis with a comprehensive theoretical and econometric analysis of model validity across stylized facts and seminal laboratory experiments to provide a rigorous assessment of model adequacy. Third, as a technical innovation in the axiomatic derivation, we formally distinguish contexts. This allows us to formalize the notion of narrow bracketing as a property of preferences, and thus to establish a formally tight but ex-ante unsuspected link between four large literatures in behavioral economics: prospect theory (Kahneman and Tversky, 1979), narrow bracketing (Read et al., 1999), altruism (Andreoni and Miller, 2002), and reference dependence (Kőszegi and Rabin, 2006). Finally, our results reconcile a wide range of seemingly inconsistent experimental results with approaches and results from classical decision theory.

By deriving a utility representation from established behavioral principles such as scaling invariance and narrow bracketing, the axiomatic approach suggests applicability of our model that goes beyond the variety of distribution games analyzed in the paper. The theoretical predictions of behavior in these games, the tight relations to four major branches of behavioral economics, and the fact that fairness-based altruism directly formalizes the widespread notion that altruism is a concern for the well-being of others, while being derived from universal behavioral axioms not specific to altruism or distribution games, renders it a promising model for future work. Our econometric results on out-of-sample adequacy provide substantial validity in this respect, and both the model's generality and its quantitative adequacy open up a number of exciting avenues for future research.

These include experimental analyses of preferences and reference points, based on an axiomatically solid and econometrically validated model, theoretical analyses of utility representations under alternative axioms and of revealed preference with non-convexities (see also Halevy et al., 2017), empirical and theoretical analyses of behavioral welfare and preference laundering,[15] and, exploiting the relation to choice under risk, behavioral analyses of giving under incomplete information (as in Dana et al., 2007, and Andreoni and Bernheim, 2009) or in multilateral interactions. Due to the large extent of similarity of charitable giving and dictator behavior in the laboratory (Konow, 2010; Huck and Rasul, 2011; DellaVigna et al., 2012), a particularly immediate range of applications lies in structural analyses of charitable giving (DellaVigna, 2009; Card et al., 2011) generalizing, for example, the work of DellaVigna et al. (2012, 2016) and Huck et al. (2015).

References

Aczél, J. (1966). *Lectures on functional equations and their applications*, volume 19. Academic press.

Almås, I., Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2010). Fairness and the development of inequality acceptance. *Science*, 328(5982):1176–1178.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal*, 100(401):464–477.

Andreoni, J. (1995). Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics*, 110(1):1–21.

[15] Letting all agents have equal weight, our analysis establishes a utilitarian welfare function which contains Rawls and Harsanyi as special cases (for $\beta \to -\infty$ and $\beta = 1$, respectively), where individual welfares are the prospect-theoretic utilities from single-person decision making. This provides an axiomatic foundation for preference laundering in welfare analyses (Goodin, 1986), i.e. to disregard concerns for others (such as envy) in behavioral welfare economics, which drastically affects policy recommendations (see also Piacquadio, 2017).

Andreoni, J. and Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636.

Andreoni, J. and Miller, J. (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753.

Andreoni, J., Rao, J. M., and Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, 125(3):625–653.

Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? image motivation and monetary incentives in behaving prosocially. *American Economic Review*, 99(1):544–555.

Auger, A., Hansen, N., Perez Zerpa, J., Ros, R., and Schoenauer, M. (2009). Experimental comparisons of derivative free optimization algorithms. *Experimental Algorithms*, pages 3–15.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11(2):122–133.

Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy*, 82(6):1063–1093.

Bellemare, C., Kröger, S., and van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, 76(4):815–839.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: Estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Blanco, M., Engelmann, D., and Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2):321–338.

Bolton, G. E. and Katok, E. (1998). An experimental test of the crowding out hypothesis: The nature of beneficent behavior. *Journal of Economic Behavior & Organization*, 37(3):315–331.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, pages 166–193.

Breitmoser, Y. (2013). Estimation of social preferences in generalized dictator games. *Economics Letters*, 121(2):192–197.

Breitmoser, Y. (2017). Discrete choice with representation effects. CRC TRR 190 Working Paper.

Broberg, T., Ellingsen, T., and Johannesson, M. (2007). Is generosity involuntary? *Economics Letters*, 94(1):32–37.

Bruhin, A., Fehr, E., and Schunk, D. (2018). The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association (forthcoming)*.

Camerer, C., Cohen, J., Fehr, E., Glimcher, P., and Laibson, D. (2017). Neuroeconomics. In Kagel, J. and Roth, A., editors, *Handbook of Experimental Economics*, volume 2, pages 153–217. Princeton University Press.

Camerer, C. and Ho, T.-H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119(3):861–898.

Cappelen, A. W., Halvorsen, T., Sørensen, E. Ø., and Tungodden, B. (2017). Face-saving or fair-minded: What motivates moral behavior? *Journal of the European Economic Association*, 15(3):540–557.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., and Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, 97(3):818–827.

Cappelen, A. W., Moene, K. O., Sørensen, E. Ø., and Tungodden, B. (2013a). Needs versus entitlements—an international fairness experiment. *Journal of the European Economic Association*, 11(3):574–598.

Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., and Tyran, J.-R. (2013b). Give and take in dictator games. *Economics Letters*, 118(2):280 – 283.

Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2010). Responsibility for what? fairness and individual responsibility. *European Economic Review*, 54(3):429–441.

Card, D., DellaVigna, S., and Malmendier, U. (2011). The role of theory in field experiments. *The Journal of Economic Perspectives*, 25(3):39–62.

Carpenter, J., Verhoogen, E., and Burks, S. (2005). The effect of stakes in distribution experiments. *Economics Letters*, 86(3):393–398.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

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

Chernoff, H. (1954). Rational selection of decision functions. *Econometrica: journal of the Econometric Society*, pages 422–443.

Cherry, T. L. (2001). Mental accounting and other-regarding behavior: Evidence from the lab. *Journal of Economic Psychology*, 22(5):605–615.

Cherry, T. L., Frykblom, P., and Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, 92(4):1218–1221.

Cherry, T. L. and Shogren, J. F. (2008). Self-interest, sympathy and the origin of endowments. *Economics Letters*, 101(1):69–72.

Cooper, D. J. and Dutcher, E. G. (2011). The dynamics of responder behavior in ultimatum games: a meta-study. *Experimental Economics*, 14(4):519–546.

Cox, J. C., Friedman, D., and Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, 59(1):17–45.

Cox, J. C., List, J. A., Price, M., Sadiraj, V., and Samek, A. (2016). Moral costs and rational choice: Theory and experimental evidence. Technical report, National Bureau of Economic Research.

Dana, J., Cain, D. M., and Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100(2):193–201.

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

De Bruyn, A. and Bolton, G. E. (2008). Estimating the influence of fairness on bargaining behavior. *Management Science*, 54(10):1774–1791.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2):315–372.

DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics*, 127:1–56.

DellaVigna, S., List, J. A., Malmendier, U., and Rao, G. (2016). Estimating social preferences and gift exchange at work. NBER Working Paper No. 22043.

Dueck, D. and Frey, B. J. (2007). Non-metric affinity propagation for unsupervised image categorization. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.

Dufwenberg, M., Heidhues, P., Kirchsteiger, G., Riedel, F., and Sobel, J. (2011). Other-regarding preferences in general equilibrium. *The Review of Economic Studies*, 78(2):613–639.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and economic behavior*, 47(2):268–298.

Easterlin, R. A. (2001). Income and happiness: Towards a unified theory. *Economic Journal*, 111(473):465–484.

Eckel, C. C., Grossman, P. J., and Johnston, R. M. (2005). An experimental test of the crowding out hypothesis. *Journal of Public Economics*, 89(8):1543–1560.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117.

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

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D. B., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, doi: 10.1093/qje/qjy013:1–48.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, pages 817–868.

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

Fleurbaey, M. and Maniquet, F. (2011). *A theory of fairness and social welfare*, volume 48. Cambridge University Press.

Gächter, S., Herrmann, B., and Thöni, C. (2004). Trust, voluntary cooperation, and socio-economic background: survey and experimental evidence. *Journal of Economic Behavior and Organization*, 55(4):505–531.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, 1(1):60–79.

Goodin, R. E. (1986). Laundering preferences. In *Foundations of social choice theory*. Cambridge University Press Cambridge.

Gul, F. and Pesendorfer, W. (2001). Temptation and self-control. *Econometrica*, 69(6):1403–1435.

Halevy, Y., Persitz, D., Zrill, L., et al. (2017). Parametric recoverability of preferences. *Journal of Political Economy*.

Harless, D. W., Camerer, C. F., et al. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, 62(6):1251–1289.

Harrison, G. W. and Johnson, L. T. (2006). Identifying altruism in the laboratory. In *Experiments Investigating Fundraising and Charitable Contributors*, pages 177–223. Emerald Group Publishing Limited.

Hey, J. D., Lotito, G., and Maffioletti, A. (2010). The descriptive and predictive adequacy of theories of decision making under uncertainty/ambiguity. *Journal of Risk and Uncertainty*, 41(2):81–111.

Hoffman, E., McCabe, K., Shachat, K., and Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7(3):346–380.

Hoffman, E., McCabe, K., and Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *American Economic Review*, 86(3):653–660.

Huck, S. and Rasul, I. (2011). Matched fundraising: Evidence from a natural field experiment. *Journal of Public Economics*, 95(5):351–362.

Huck, S., Rasul, I., and Shephard, A. (2015). Comparing charitable fundraising schemes: Evidence from a natural field experiment and a structural model. *American Economic Journal: Economic Policy*, 7(2):326–69.

Jakiela, P. (2011). Social preferences and fairness norms as informal institutions: experimental evidence. *American Economic Review*, 101(3):509–513.

Jakiela, P. (2015). How fair shares compare: Experimental evidence from two cultures. *Journal of Economic Behavior & Organization*, 118:40–54.

Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5):865–889.

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, pages 728–741.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, 5(1):193–206.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292.

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

Konow, J. (2003). Which is the fairest one of all? a positive analysis of justice theories. *Journal of Economic Literature*, 41(4):1188–1239.

Konow, J. (2010). Mixed feelings: Theories of and evidence on giving. *Journal of Public Economics*, 94(3):279–297.

Korenok, O., Millner, E. L., and Razzolini, L. (2012). Are dictators averse to inequality? *Journal of Economic Behavior & Organization*, 82(2):543–547.

Korenok, O., Millner, E. L., and Razzolini, L. (2013). Impure altruism in dictators' giving. *Journal of Public Economics*, 97:1–8.

Korenok, O., Millner, E. L., and Razzolini, L. (2014). Taking, giving, and impure altruism in dictator games. *Experimental Economics*, 17(3):488–500.

Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Kőszegi, B. and Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073.

Kőszegi, B. and Rabin, M. (2009). Reference-dependent consumption plans. *American Economic Review*, 99(3):909–936.

Kritikos, A. and Bolle, F. (2005). Utility-based altruism: evidence from experiments. In *Psychology, Rationality and Economic Behaviour*, pages 181–194. Springer.

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

Lazear, E. P., Malmendier, U., and Weber, R. A. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics*, 4(1):136–163.

List, J. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493.

List, J. (2009). Social preferences: Some thoughts from the field. *Annual Review of Economics*, 1(1):563–583.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. John Wiley and sons.

Malmendier, U., te Velde, V. L., and Weber, R. A. (2014). Rethinking reciprocity. *Annual Review of Economics*, 6(1):849–874.

McCullough, B. D. and Vinod, H. D. (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, 93(3):873–892.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Oosterbeek, H., Sloof, R., and Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2):171–188.

Oxoby, R. J. and Spraggon, J. (2008). Mine and yours: Property rights in dictator games. *Journal of Economic Behavior & Organization*, 65(3):703–713.

Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *The Journal of Neuroscience*, 29(44):14004–14014.

Padoa-Schioppa, C. and Rustichini, A. (2014). Rational attention and adaptive coding: a puzzle and a solution. *American Economic Review*, 104(5):507–513.

Piacquadio, P. G. (2017). A fairness justification of utilitarianism. *Econometrica*, 85(4):1261–1276.

Powell, M. (2006). The newuoa software for unconstrained optimization without derivatives. *Large-Scale Nonlinear Optimization*, pages 255–297.

Rabin, M. and Weizsäcker, G. (2009). Narrow bracketing and dominated choices. *American Economic Review*, 99(4):1508–1543.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Read, D., Loewenstein, G., and Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, 19(1-3):171–197.

Rohde, K. I. (2010). A preference foundation for fehr and schmidt's model of inequity aversion. *Social Choice and Welfare*, 34(4):537–547.

Rubinstein, A. (2012). *Lecture notes in microeconomic theory: the economic agent*. Princeton University Press.

Ruffle, B. J. (1998). More is better, but fair is fair: Tipping in dictator and ultimatum games. *Games and Economic Behavior*, 23(2):247–265.

Saito, K. (2013). Social preferences under risk: equality of opportunity versus equality of outcome. *American Economic Review*, 103(7):3084–3101.

Samuelson, W. and Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1):7–59.

Schennach, S. and Wilhelm, D. (2016). A simple parametric model selection test. *Journal of the American Statistical Association*.

Schmidt, U. (2003). Reference dependence in cumulative prospect theory. *Journal of Mathematical Psychology*, 47(2):122–131.

Sen, A. K. (1971). Choice functions and revealed preference. *The Review of Economic Studies*, 38(3):307–317.

Simonsohn, U. and Gino, F. (2013). Daily horizons: evidence of narrow bracketing in judgment from 10 years of mba admissions interviews. *Psychological science*, 24(2):219–224.

Skiadas, C. (2013). Scale-invariant uncertainty-averse preferences and source-dependent constant relative risk aversion. *Theoretical Economics*, 8(1):59–93.

Skiadas, C. (2016). Scale or translation invariant additive preferences. Unpublished manuscript.

Tremblay, L. and Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729):704–708.

Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics*, pages 1039–1061.

van der Weele, J. J., Kulisa, J., Kosfeld, M., and Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *American Economic Journal: Microeconomics*, 6(3):256–264.

Wakker, P. and Tversky, A. (1993). An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty*, 7(2):147–175.

Wakker, P. P. (1989). *Additive representations of preferences: A new foundation of decision analysis*, volume 4. Springer Science & Business Media.

Wakker, P. P. and Zank, H. (2002). A simple preference foundation of cumulative prospect theory with power utility. *European Economic Review*, 46(7):1253–1271.

Wilcox, N. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In Cox, J. C. and Harrison, G. W., editors, *Risk aversion in experiments*, volume 12 of *Research in experimental economics*, pages 197–292. Emerald Group Publishing Limited.

Wilcox, N. T. (2011). Stochastically more risk averse: A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, 162(1):89–104.

Wilcox, N. T. (2015). Error and generalization in discrete choice under risk. ESI Working Paper 15-11.

Appendix (For online publication)

Fairness-based altruism

Yves Breitmoser and Pauline Vorjohann
Bielefeld University and University of Exeter

A Relation to social norms and "social appropriateness"

Starting with Krupka and Weber (2013), a growing literature relates giving observed in experiments to norm compliance. Subjects are assumed to have a common understanding of the "social appropriateness" of options, which in turn affects dictator behavior and is a function of the social norms applying in a given context. In a novel experimental design, Krupka and Weber measure social appropriateness by having (third) subjects play a coordination game—asking each subject how "socially appropriate" the available options are in the eyes of their co-players and paying a prize to all subjects picking the modal response. The mean of all appropriateness ratings is mapped into a measure $s_x \in [-1,1]$ for all options x, with $s_x = -1$ indicating highly inappropriate and $s_x = 1$ indicating highly appropriate options. Krupka and Weber then examine if a utility function of the form

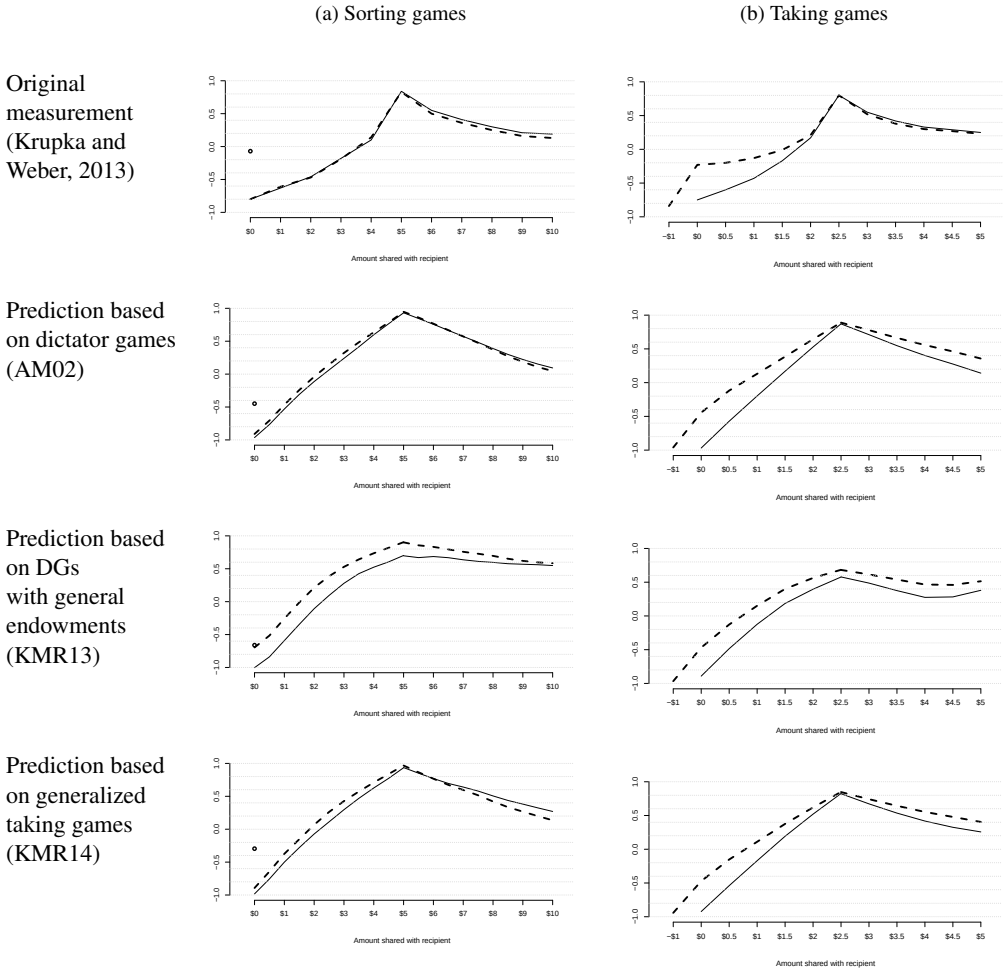
$$u_x = \pi_x + \alpha s_x \tag{4}$$

fits behavior observed in earlier dictator game experiments, using the weight α as a free parameter. While statistical tests supporting the results are not provided, the plots in Krupka and Weber (2013) suggest a good fit after calibrating α. This finding has been interpreted as indicating that behavior is norm-guided, rather than being payoff or welfare concerned as assumed in earlier work. In the following, we clarify the relation of our findings to those of Krupka and Weber (2013) and subsequent work, to discuss how we may think of fairness-based altruism as a foundation of norm-guided giving.

To this end, let us recap two main results. Krupka and Weber convincingly demonstrate that experimental subjects are able to predict behavior in taking and sorting games, a feat that existing behavioral models struggled to achieve. We have shown that fairness-based altruism also allows to predict behavior, and hence our conjecture: the two are likely to correlate. A post-hoc straightforward approach would be to take our predictions of utility u_x across options, the respectively induced payoffs π_x, and to then compute social appropriateness s_x by inverting Eq. (4) for all options x. We skip this fairly unintelligible exercise and evaluate whether social appropriateness may be deduced from first principles.

Krupka and Weber (2013) interpret social appropriateness as reflecting the social norm that dictators facing a specific dictator game trade off with their self-interest. They argue that since their elicitation method (i) makes uninvolved subjects rate actions rather than outcomes and (ii) incentivizes subjects to rate in accordance with what they regard as a socially shared assessment, the resulting appropriateness ratings satisfy the two main characteristics of a social norm as defined by Elster (1989). These defining features of social norms are closely related to the "social contract" of Rawls (1971), which specifies a standard for social and distributive justice that "free and rational persons concerned to further their own interests would accept in an initial position of equality" (p. 11). The idea is that the members of a society would unanimously agree to the

Figure 3: Relation of experimentally measured "social appropriateness" (Krupka and Weber) to the Rawlsian prediction following from our estimates



(c) Correlation between observed and predicted appropriateness

	Sorting games		Taking games	
Predictions based on …	Spearman-ρ	p-value	Spearman-ρ	p-value
Dictator games (AM02)	0.641	(0.001)	0.738	(0)
Gen endowments (KMR13)	0.667	(0.001)	0.766	(0)
Taking games (KMR14)	0.644	(0.001)	0.751	(0)

Note: The "sorting games" compare appropriateness in a standard dictator game with endowments of 10 for the dictator and 0 for the recipient to appropriateness in a sorting game where the dictator game is succeeded by giving the dictator the option to sort out at costs of 1. The "taking games" compare appropriateness in a standard dictator game with endowments of 10 for the dictator and 5 for the recipient to appropriateness in a taking game where the dictator game may alternatively take one currency unit from the recipient's endowment. The plots follow Krupka and Weber: solid lines represent the social appropriateness in the standard dictator games and dashed lines represent social appropriateness in the sorting and taking games, respectively. The single "dot" in the sorting games reflects the appropriateness of sorting out.

social contract if they met behind the "veil of ignorance", a hypothetical place where they are unaware of their positions in society (see also Konow, 2003). According to Rawls (1971) the social contract emerging in such a situation would prescribe a distribution that equalizes individual welfares unless inequality is to the advantage of the individual with the minimum welfare. While for obvious reasons an experimental test of Rawls hypothesis can never be perfect, Krupka and Weber's subjects share some central characteristics with Rawls' society members behind the veil of ignorance. They can be thought of as impartial since they are uninvolved while they are part of the same society as the involved players. Furthermore, they are incentivized to find an agreement instead of simply voicing their opinions. Therefore, looking at Krupka and Weber's social appropriateness ratings through the lense of our fairness-based altruism model allows us to test the Rawlsian hypothesis of social welfare being the minimum of all individual welfares, joint with the assertion that social appropriateness simply transforms social welfare to a scale ranging from highly inappropriate (-1) to highly appropriate (1).

Since our fairness-based approach directly builds on individual welfares v_1 and v_2, we are able to directly test the asserted Rawlsian link between appropriateness and welfares—simply by predicting individual welfares for all options in the sorting and taking games analyzed by Krupka and Weber, taking the minimum of v_1 and v_2 across options, and rescaling such that a measure ranging from -1 to $+1$ results. Specifically, we predict the social appropriateness ratings for both taking and sorting games analyzed by Krupka and Weber based on our estimates from each of the three experiments analyzed before (AM02, MKR13, and KMR14). This yields 3×2 profiles of appropriateness ratings, which we then relate to the measurements of Krupka and Weber.[16] The results are reported in Figure 3 and strongly corroborate the relation of social appropriateness and Rawlsian welfare asserted already by Krupka and Weber. The correlation between the out-of-sample predictions and the in-sample measurements of Krupka and Weber is very high, around 0.65 in sorting games and around 0.75 in taking games, regardless of the data set which the prediction is based on. We therefore conclude as follows.

Result 3. *Krupka and Weber's measure of social appropriateness strongly correlates with the Rawlsian notion of welfare, based on out-of-sample predictions of individual welfares derived from the above model of fairness-based altruism.*

That is, social appropriateness is founded in welfare concerns in the intuitive Rawlsian manner alluded to by Krupka and Weber. It seems futile to ask which came first, welfare concerns or social appropriateness/social norms, they rather appear to be two sides of the same coin. The received interpretation that giving reflects context-dependent social norms rather than more fundamental payoff and welfare concerns seems premature, but so would the opposite. From a practical point of view, both approaches seem to have distinctive strengths. Analyses relating behavior to social appropriateness need not be concerned with individual preferences and can focus on the picture at large. In turn, the behavioral foundation in welfare concerns has an independent axiomatic foundation in established behavioral principles, which greatly facilitates application across contexts, and the implied S-shape of individual welfares has been observed in many contexts, which promises reliable predictions and policy recommendations out-of-sample.

[16] Specifically, for each subject in our in-sample experiments (AM02, KMR13, KMR14), we determine the individual welfares if that subject would play either role, v_1 and v_2. We then assume that an impartial observer in the sense of Krupka and Weber determines appropriateness as follows: Across dictators, what is their average individual welfare from choosing x conditional on choosing x in the first place. Across recipients, what is their average individual welfare from getting x conditionally on being confronted with x in the first place (which is an empty condition, stated only for symmetry). The lesser of these conditional expectations is the unscaled Rawlsian appropriateness of each option, and rescaling to $[-1, 1]$ across options yields our out-of-sample prediction for Krupka-Weber appropriateness.

B Relegated proofs

B.1 Proof of Proposition 1

The proof is provided by a sequence of lemmas, and in all lemmas, we maintain Assumption 1. Initially, we establish $\Leftarrow$, i.e. that the claimed representations satisfy the respective axioms. Afterwards, we demonstrate that the axioms imply the claimed representation.

Lemma 1 ($\Leftarrow$, Part 1)**.** *For all $\pi \in \Pi$, if $\succsim_\pi$ is represented by a continuous u_π satisfying $u_\pi(x') = \sum_{i\leq n} \alpha_i \cdot v_i\big[\pi_i(x')\big]$ for all x', with*

$$v_i(p) \underset{\beta\neq 0}{=} \begin{cases} p^\beta/\beta, & \text{if } p \geq 0 \\ -\delta_i \cdot (-p)^\beta/\beta, & \text{if } p < 0 \end{cases} \qquad \text{and} \qquad v_i(p) \underset{\beta=0}{=} \log(p)$$

for some $\alpha \in \mathbb{R}^n$, $\beta \in \mathbb{R}$, $\delta \in \mathbb{R}^n$, then $\succsim_\pi$ satisfies Axioms (1)–(4),(5).

Proof. Axioms (1)–(3) are immediate, see also Theorem III.4.1 in Wakker (1989). Axiom (5) follows from the context independence of the parameters (α, β, δ). Finally, any context is scaling invariant in the sense of Axiom (4). To see this, note that scaling the outcome vector $\pi(x)$ induces either linear transformations or translations of utilities and therefore does not affect the preference ordering. □

Lemma 2 ($\Leftarrow$, Part 2)**.** *If there exist $\alpha \in \mathbb{R}^n$, $\beta \in \mathbb{R}$, $\delta \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^{n\times n}$ such that for all contexts $\pi \in \Pi$, $\succsim_\pi$ is represented by u_π satisfying $u_\pi(x') = \sum_{i\leq n} \alpha_i \cdot v_i\big[\pi_i(x') - r_i(\pi)\big]$ for all x', with*

$$v_i(p) \underset{\beta\neq 0}{=} \begin{cases} p^\beta/\beta, & \text{if } p \geq 0 \\ -\delta_{\pi,i} \cdot (-p)^\beta/\beta, & \text{if } p < 0 \end{cases} \qquad \text{and} \qquad v_i(p) \underset{\beta=0}{=} \log(p)$$

and

$$r_i = \min \pi_i + \sum_{j\neq n} w_{i,j} \cdot (\pi_j(0) - \min \pi_j),$$

then $\succsim_\pi$ satisfies Axioms (1)–(3), (4), (6), (7)–(8) for all $\pi \in \Pi$.

Proof. Axioms (1)–(3) and (8) are immediate. Scaling invariance in the sense of Axiom (4) follows from the observation that any context π' satisfying $r(\pi') = \mathbf{0}$ is scaling invariant, and that for any context π, the context $\pi' = \pi - r(\pi)$ (which exists by context richness) satisfies both $r(\pi') = \mathbf{0}$ and $\pi'(0) - \min \pi' = \pi(0) - \min \pi$. Regarding narrow bracketing in the sense of Axiom (6), observe that under the above representation, for any $\pi, \pi' \in \Pi$, $\pi(0) - \min \pi = \pi'(0) - \min \pi'$ implies $r(\pi) = r(\pi')$ and thus equivalence of the utility representation in the sense of Axiom (6).

Finally, consider (additive) compensability, i.e. Axiom (7). On one hand, consider any $\pi, \pi' \in \Pi$. We seek to demonstrate that there exists $c \in \mathbb{R}^n$ such that $\pi' \sim \pi + c$. Given the equality of (α, β, δ), this obtains for $c = r(\pi') - r(\pi)$, which is well-defined for all $\mathbf{w}$. For, given any x, y, x', y' such that $\pi(x) = \pi'(x')$ and $\pi(y) = \pi'(y')$, we have

$$\begin{aligned}
&(\pi + c)(x) \succsim_{\pi+c} (\pi + c)(y) \\
&\quad \Leftrightarrow \quad \sum_{i\leq n} \alpha_i \cdot v_i\big[(\pi + c)_i(x) - r_i(\pi + c)\big] \geq \sum_{i\leq n} \alpha_i \cdot v_i\big[(\pi + c)_i(y) - r_i(\pi + c)\big] \\
&\quad \Leftrightarrow \quad \sum_{i\leq n} \alpha_i \cdot v_i\big[\pi_i(x) + c_i - r_i(\pi) - c_i\big] \geq \sum_{i\leq n} \alpha_i \cdot v_i\big[\pi_i(y) + c_i - r_i(\pi) - c_i\big] \\
&\quad \Leftrightarrow \quad \sum_{i\leq n} \alpha_i \cdot v_i\big[\pi'_i(x') - r_i(\pi')\big] \geq \sum_{i\leq n} \alpha_i \cdot v_i\big[\pi'_i(y') - r_i(\pi')\big] \qquad \Leftrightarrow \qquad \pi'(x') \succsim_{\pi'} \pi'(y'),
\end{aligned}$$

since $\pi'_i(x') = \pi_i(x) + c_i$ and $r_i(\pi') = r_i(\pi) + c_i$ for all i by construction.

One the other hand, we have to show that the compensation is additive, i.e. that $\pi' \sim \pi + c'$ and $\pi'' \sim \pi + c''$ implies $\pi' + \pi'' \sim 2\pi + c' + c''$. Given the previous observation, that $c = r(\pi') - r(\pi)$ is the compensation, we have to show that

$$r(\pi') = r(\pi) + c' \text{ and } r(\pi'') = r(\pi) + c'' \quad \Longrightarrow \quad r(\pi' + \pi'') = r(2\pi) + c' + c''.$$

This obtains, as $r(\pi' + \pi'') = r(\pi') + r(\pi'')$ and $r(2\pi) = 2r(\pi)$ follows by linearity of r. Context continuity, finally, obtains by continuity of (v_i) and r. □

Lemma 3 (Beginning of $\Rightarrow$, Separable utility function)**.** *For each $\pi \in \Pi$, given Axioms (1)–(3) there exists a continuous $v_\pi : \mathbb{R}^n \to \mathbb{R}^n$ such that*

$$u_\pi(x') = \sum_{i \leq n} v_{\pi,i}\big(\pi_i(x')\big) \tag{5}$$

represents $\succsim_\pi$.

Proof. Axioms (1)–(2) imply existence of a continuous utility representation (see e.g. Rubinstein, 2012, chap. 4). In addition Axiom (3) implies existence of an additively separable utility representation, see Theorem III.4.1 in Wakker (1989) for each context $\pi \in \Pi$. That is, there exists a family of functions $\{v_{\pi,i} : \mathbb{R} \to \mathbb{R}\}_{\pi \in \Pi, i \leq n}$ such that $\pi(x) \succsim_\pi \pi(y) \Leftrightarrow u_\pi(x) \geq u_\pi(y)$ for all $x, y \in X$ and $\pi \in \Pi$ with

$$u_\pi(x') = \sum_{i \leq n} v_{\pi,i}\big(\pi_i(x')\big) \tag{6}$$

for all $x' \in X, \pi \in \Pi$. For later reference, Wakker's Theorem III.4.1 also establishes that all additively separable representations $\tilde{u}_\pi$ of $\succsim_\pi$ are positive affine transformations of one another. Also note that the representations obtained so far may be context dependent. □

Lemma 4 (Context independence by broad bracketing)**.** *Given Axioms (1)–(3), (5), there exists $v : \mathbb{R}^n \to \mathbb{R}^n$ such that, for all $\pi \in \Pi$,*

$$u_\pi(x) = \sum_{i \leq n} v_i\big(\pi_i(x)\big) \tag{7}$$

represents $\succsim_\pi$.

Proof. For each context π, fix value functions $(v_{\pi,i})$ representing $\succsim_\pi$ as existent by Lemma 3.

Fix any context $\pi \in \Pi$. Since $\pi + c \in \Pi$ for all $c \in \mathbb{R}^n$, the union of the images $\cup_{c \in \mathbb{R}^n}(\pi + c)[X]$ is equal to $\mathbb{R}^n$. Since any $\pi[X] - \min \pi$ is a non-degenerate cone in $\mathbb{R}^n$, we can also define a countable subset $C \subset \mathbb{R}^n$ such that $\cup_{c \in C}(\pi + c)[X] = \mathbb{R}^n$, take for example $\mathbb{Q}^n$. Given this, we can construct the function $v : \mathbb{R}^n \to \mathbb{R}^n$ claimed to exist by induction over C. Having fixed π above, let $o : \mathbb{N} \to C$ denote any linear ordering over C with $o(1) = \mathbf{0}$ such that $v_{\pi + o(k)}[X] \cap v_{\pi + o(k+1)}[X]$ has positive measure in $\mathbb{R}^n$, for all $k \geq 0$. We start by defining $v_1(p) := v_\pi(p)$ for all $p \in \pi[X]$, noting that $\pi = \pi + o(1)$. Next, fix $k > 1$, let $\pi' = \pi + o(k)$ and let $P_k = \cup_{k' < k}(\pi + o(k'))[X]$ denote the set of points "previously defined". Given this, let $P' = \pi'[X] \cap P_k$ denote the overlap to points previously defined. Now, pick a transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ such that for all x, x' and all $i \leq n$, "monotonicity" obtains as follows

$$v_i \geq v_i' \Rightarrow f_i(v_i) \geq f(v_i') \quad \forall i \qquad \text{and} \qquad \sum_i v_i \geq \sum_i v_i' \Rightarrow \sum_i f_i(v_i) \geq \sum_i f(v_i'),$$

while $f_i(v_{\pi',i}(p)) = v_{k,i}(p)$ for all $p \in P'$ and $i \leq n$. Such a transformation exists by broad bracketing (indeed, an affine one exists). Then, define $v_{k,i}(p') = f_i(v_{\pi',i}(p'))$ for all $p' \in \pi'[X] \setminus P'$, i.e. for the new points. Succesively, we thus define $v_k(x)$ for all $x \in \mathbb{R}^n$, as k increases.

Let v be the pointwise limit of v_k as k tends to infinity, which is well-defined by construction for the entire domain $\mathbb{R}^n$. It remains to show that

$$u_\pi(x) = \sum_{i\leq n} v_i\big(\pi_i(x)\big) \tag{8}$$

represents $\succsim_\pi$ for all π. For contradiction, assume the opposite, i.e. that there exist x, x', π' such that

$$\pi'(x) \succsim_{\pi'} \pi'(x') \qquad \text{and} \qquad \sum_{i\leq n} v_i\big(\pi'_i(x)\big) < \sum_{i\leq n} v_i\big(\pi'_i(x')\big).$$

Let k and k' denote the induction steps in which $\pi'(x)$ and $\pi'(x')$ had been added to the domain of v, i.e. k such that $\pi(x) \in P_k \setminus P_{k-1}$, using $P_0 = \emptyset$, and k' such that $\pi(x') \in P_{k'} \setminus P_{k'-1}$. If $k = k'$ and $\pi' = \pi + o(k)$, as defined above, then it directly violates "monotonicity" of the functions f_i as defined above. If $k = k'$ and $\pi' \neq \pi + o(k)$, as defined above, then it either violates monotonicity again, if for $\pi'' = \pi + o(k)$,

$$\pi'(x) \succsim_{\pi''} \pi'(x') \qquad \text{and} \qquad \sum_{i\leq n} v_i\big(\pi'_i(x)\big) < \sum_{i\leq n} v_i\big(\pi'_i(x')\big),$$

or broad bracketing, as

$$\pi'(x) \succsim_{\pi'} \pi'(x') \qquad \text{and} \qquad \pi'(x) \not\succsim_{\pi''} \pi'(x').$$

Finally, if $k \neq k'$, we again obtain a violation of montonicity, as

$$\sum_{i\leq n} v_{\pi',i}\big(\pi'_i(x)\big) \geq \sum_{i\leq n} v_{\pi',i}\big(\pi'_{\pi',i}(x')\big) \qquad \text{and} \qquad \sum_{i\leq n} v_i\big(\pi'_i(x)\big) < \sum_{i\leq n} v_i\big(\pi'_i(x')\big).$$

□

Lemma 5 (Reference dependence by narrow bracketing)**.** *Given Axioms (1)–(3), (4), (6), there exists a family of functions* $\{v_\pi : \mathbb{R}^n \to \mathbb{R}^n\}_{\pi\in\Pi}$ *and* $f : \mathbb{R}^n \to \mathbb{R}^n$ *such that*

$$u_\pi(x) = \sum_{i\leq n} v_{\pi,i}\big(\pi_i(x) - r_i(\pi)\big) \qquad \textit{and} \qquad r(\pi) = \min\pi + f(\pi(0) - \min\pi) \tag{9}$$

represents $\succsim_\pi$ *for all* $\pi \in \Pi$*, where*

1. $v_\pi = v_{\pi'}$ *for any two* $\pi, \pi' \in \Pi$ *satisfying* $\pi(0) - \min\pi = \pi'(0) - \min\pi'$*,*
2. $r(\pi) = \min\pi - \min\pi^0$ *using* $\pi^0 = S^0(\pi(0) - \min\pi)$*.*

Proof. By Axiom (4), for each context $\pi \in \Pi$ there exists a scaling invariant π^0 such that $\pi^0(0) - \min\pi^0 = \pi(0) - \min\pi$ (there may be several such contexts, but this is irrelevant for us). Hence, we can define a function $S^0 : \mathbb{R}^n \to \Pi$ such that $S^0(\pi(0) - \min\pi)$ is a scaling invariant context π^0 satisfying $\pi^0(0) - \min\pi^0 = \pi(0) - \min\pi$. Further, let Π^0 denote the set of *these* scaling invariant contexts, i.e. $\Pi^0 = \{S^0(\pi(0) - \min\pi) | \pi \in \Pi\}$.

Given this function S^0, we show that if Axioms (1)-(3) and (6) hold, then the preferences admit the claimed representation for some $r : \Pi \to \mathbb{R}^n$. Fix this r and any $\pi', \pi' \in \Pi$ such that $\pi'(0) - \min\pi' = \pi(0) - \min\pi$. By narrow bracketing and the utility representations obtained in Lemma 3, there exists a function $\tilde{v} : \mathbb{R}^n \to \mathbb{R}$ such that

$$\sum_{i\leq n} \tilde{v}\big(\pi_i(x) - \min\pi\big) \geq \sum_{i\leq n} \tilde{v}\big(\pi_i(y) - \min\pi\big) \\ \Leftrightarrow \quad \sum_{i\leq n} \tilde{v}\big(\pi'_i(x) - \min\pi'\big) \geq \sum_{i\leq n} \tilde{v}\big(\pi'_i(y) - \min\pi'\big) \tag{10}$$

for all $x,y \in X$. Using $r(\pi') = \min \pi'$ for all π', this implies that that there exists a function $\tilde{v}$: $\mathbb{R}^n \to \mathbb{R}$ such that for all $\pi' \in \Pi$ satisfying $\pi'(0) - \min \pi' = \pi(0) - \min \pi$ and all $x,y \in X$,

$$\sum_{i \leq n} \tilde{v}\big(\pi_i(x) - r_i(\pi)\big) \geq \sum_{i \leq n} \tilde{v}\big(\pi_i(y) - r_i(\pi)\big) \quad \Leftrightarrow \quad \sum_{i \leq n} \tilde{v}\big(\pi'_i(x) - r_i(\pi')\big) \geq \sum_{i \leq n} \tilde{v}\big(\pi'_i(y) - r_i(\pi')\big).$$

Since this holds true for all $\pi' \in \Pi$, the claim is established using $v_i = \tilde{v}_i$ for all $i \leq n$. Note again that u (and thus v) is unique up to positive affine transformation.

Finally, fix the scaling-invariant context $\pi^0 = S^0(\pi)$, which exists by Axiom (4), and note that

$$u_{\pi^0}(x) = \sum_{i \leq n} v_i\big(\pi_i^0(x) - r_i(\pi^0)\big), \tag{11}$$

implies that we can translate (v_i) and (r_i) such that

$$u_{\pi^0}(x) = \sum_{i \leq n} v_i\big(\pi_i^0(x)\big), \tag{12}$$

i.e. such that $r_i(\pi^0) = 0$ for all $i \leq n$. As a result of this translation, $r(\pi) = \min \pi - \min \pi^0$ using $\pi^0 = S^0(\pi(0) - \min \pi)$ for all π. Note that given this translation, we can analyze narrow bracketing and broad bracketing in a uniform manner when focusing on $\pi^0 \in \Pi^0$ (i.e. we do not have to include r_i as $r_i(\pi^0) = 0$). Finally, also note that we obtain $r(\pi) = \min \pi + c$ with $c = \min S^0(\pi(0) - \min \pi)$, or as claimed, $r(\pi) = \min \pi + f(\pi(0) - \min \pi)$ where $f(\pi(0) - \min \pi) = \min S^0(\pi(0) - \min \pi)$. □

Lemma 6 (Scaling invariance and characterization for $\pi \in \Pi^0$)**.** *For any $\pi \in \Pi^0$,*

$$v_{\pi,i}(p_i) = \beta_\pi \cdot \log p_i \qquad \textit{or} \qquad v_{\pi,i}(p_i) = \begin{cases} \alpha^+_{\pi,i} \cdot (p_i)^{\beta_\pi}, & \textit{if } p_i \geq 0, \\ -\alpha^-_{\pi,i} \cdot (-p_i)^{\beta_\pi}, & \textit{if } p_i < 0, \end{cases}$$

with $\alpha^+_{\pi,i}, \alpha^-_{\pi,i} \neq 0$ and $\beta_\pi \in \mathbb{R}$ for all $i \leq n$.

Proof. Fix any $\pi^0 \in \Pi^0$. By Axiom (4), respectively, preferences in any context $\pi^0 \in \Pi^0$ are scaling invariant. That is, for all $\lambda > 0$, define $u_{\lambda\pi^0} : X \to \mathbb{R}$ such that

$$u_{\lambda\pi^0}(x) = \sum_{i \leq n} v_i\big(\lambda \pi_i^0(x)\big), \tag{13}$$

for all λ, x, and we obtain

$$u_{\lambda\pi^0}(x) \geq u_{\lambda\pi^0}(y) \qquad \Leftrightarrow \qquad u_{\pi^0}(x) \geq u_{\pi^0}(y) \Leftrightarrow \qquad \pi^0(x) \succsim_{\pi^0} \pi^0(y). \tag{14}$$

That is, both $u_{\lambda\pi^0}$ and u_{π^0} are additively separable representations of $\succsim_{\pi^0}$, which implies (see Theorem III.4.1 of Wakker, 1989) that $u_{\lambda\pi^0}$ is a positive affine transformation of u_{π^0}, i.e. there exist $a : \mathbb{R}_+ \to \mathbb{R}$ and $b : \mathbb{R}_+ \to \mathbb{R}^n$ such that

$$v_i\big(\lambda \pi_i^0(x)\big) = v_i\big(\pi_i^0(x)\big) \cdot a(\lambda) + b_i(\lambda) \tag{15}$$

for all $i \in N$, $x \in X$, $\lambda \in \Lambda$. Now, define $X_i^+ = \{x \in X \,|\, \pi_i^0(x) > 0\}$ as well as $\tilde{\lambda} = \log \lambda$, $\tilde{v}_i : \mathbb{R} \to \mathbb{R}$ such that $\tilde{v}_i(\log p) = v_i(p)$ for all $p > 0$, and $\tilde{\pi}_i^0(x) = \log \pi_i^0(x)$ for all $x \in X_i^+$, which yields

$$\tilde{v}_i\big(\tilde{\lambda} + \tilde{\pi}_i^0(x)\big) = \tilde{v}_i\big(\tilde{\pi}_i^0(x)\big) \cdot a(\tilde{\lambda}) + b_i(\tilde{\lambda}). \tag{16}$$

By continuity of v_i we obtain continuity of $\tilde{v}_i$, and since the payoff image $\pi^0[X]$ is a cone in $\mathbb{R}^n$ with all dimensions being essential, it has positive volume in $\mathbb{R}^n$, i.e. $\pi_i^0[X]$ is an interval of positive length for all dimensions i. By Theorem 1 of Aczél (1966, p. 150), all solutions of this (Pexider)

functional equation satisfy (besides a constant solution that is ruled out by essentialness) either

$$\text{Case 1:} \qquad \tilde{v}_i(p) = f_i(p) + \alpha_i, \qquad a(\tilde{\lambda}) = 1, \qquad b_i(\tilde{\lambda}) = f_i(\tilde{\lambda}), \tag{17}$$

or

$$\text{Case 2:} \qquad \tilde{v}_i(p) = \gamma_i \cdot e^{f_i(p)} + \alpha_i, \qquad a(\tilde{\lambda}) = e^{f_i(\tilde{\lambda})}, \qquad b_i(\tilde{\lambda}) = \alpha_i \cdot \left[1 - e^{f_i(\tilde{\lambda})}\right], \tag{18}$$

where $\gamma \neq 0$ and α are arbitrary constants, and f_i is an arbitrary solution of Cauchy's fundamental (functional) equation

$$f_i(x+y) = f_i(x) + f_i(y). \tag{19}$$

Case 1: In this case, f_i is defined as $f_i(p) = \tilde{v}_i(p) - \tilde{v}_i(0)$. Function f_i is continuous by continuity of v_i, and defined on an interval with positive length since the payoff image is a cone, implying that its general solution is $f_i(p) = c_i p_i + a_i$ (see Theorem 1 on page 46 of Aczél, 1966). Hence, $\tilde{v}_i(p) = c_i p_i + \alpha_i$ for some $\alpha_i \in \mathbb{R}$. Changing notation and inverting the variable substitution, we obtain

$$\tilde{v}_i\big(\tilde{\pi}_i^0(x)\big) = \alpha \cdot \tilde{\pi}_i^0(x) + \gamma \qquad \Rightarrow \qquad v_i(\pi_i^0(x)) = \alpha \cdot \log \pi_i^0(x) + \gamma$$

with $\alpha \neq 0$ and γ being arbitrary constants.

Case 2: In this case, f_i characterizes the solution $a(\tilde{\lambda}) = e^{f_i(\tilde{\lambda})}$ of the Cauchy-type functional equation $a(x+y) = a(x) \cdot a(y)$ for the function a defined above. Since the payoff image is a cone, function a is defined on an interval containing $[0,1]$, and by inversion, all reciprocals, i.e. a is defined for the nonnegative reals. Hence, its general non-constant and continuous solution (see Theorem 1 on page 38 of Aczél, 1966) satisfies $a(x) = e^{cx}$ for some $c \neq 0$, i.e. $f_i(x) = e^{cx}$, and thus $\tilde{v}_i(p) = \gamma_i \cdot e^{c_i p} + \alpha_i$. Again, changing notation and inverting the variable substitution, we obtain

$$\tilde{v}_i\big(\tilde{\pi}_i^0(x)\big) = \alpha \cdot e^{\beta \tilde{\pi}_i^0(x)} + \gamma \qquad \Rightarrow \qquad v_i(\pi_i^0(x)) = \alpha \cdot \big(\pi_i^0(x)\big)^{\beta} + \gamma.$$

with $\alpha \neq 0$ and β, γ being arbitrary constants. [End of case distinction]

To distinguish the constants from constants in other dimensions, we rewrite

$$v_i(\pi_i^0(x)) = \alpha_i^+ + \beta_i^+ \cdot \log \pi_i^0(x) \qquad \text{or} \qquad v_i(\pi_i^0(x)) = \alpha_i^+ \cdot \big(\pi_i^0(x)\big)^{\beta_i^+} + \gamma_i^+$$

for all $x \in X_i^+$. Next, define $X_i^- = \{x \in X \,|\, \pi_i^0(x) < 0\}$, and if the set is not empty, apply the same line of arguments to $-\pi_i^0(x)$ for all $x \in X_i^-$, which yields

$$v_i(\pi_i^0(x)) = \alpha_i^- + \beta_i^- \cdot \log\big(-\pi_i^0(x)\big) \qquad \text{or} \qquad v_i(\pi_i^0(x)) = -\alpha_i^- \cdot \big(-\pi_i^0(x)\big)^{\beta_i^-} + \gamma_i^-$$

for all $x \in X_i^-$, again with $\alpha_i^- \neq 0$ and β_i^-, γ_i^- being arbitrary constants.

In the following, we refer to the two possible forms of the value function v_i as power form and logarithmic form (in the obvious manner). By continuity, the logarithmic form is feasible only if $\pi_i(x) > 0$ for all $x \in X$, implying that the second branch is never taken. Hence, for all $i \leq n$ and all $x \in X$,

$$v_i(\pi_i^0(x)) = \alpha_i^+ + \beta_i^+ \cdot \log\big(\pi_i^0(x)\big),$$

and we can set $\alpha_i^+ = 0$ for all i by applying a positive affine transformation (recalling that the value functions are unique up to positive affine transformation). This establishes the proposition's claim for the logarithmic form in any context π^0, noting that α_i^+ and β_i^+ are switched (for the logarithmic form) in the formulation of the proposition for notational convenience.

Regarding the power form of the value function, rescaling payoffs we obtain

$$\forall x \in X_i^+ : \; v_i(\lambda\pi_i^0(x)) = \alpha_i^+ \cdot \big(\lambda\pi_i^0(x)\big)^{\beta_i^+} + \gamma_i^+ = \alpha_i^+ \cdot \big(\pi_i^0(x)\big)^{\beta_i^+} \cdot \lambda^{\beta_i^+} + \gamma_i^+$$
$$\forall x \in X_i^- : \; v_i(\lambda\pi_i^0(x)) = -\alpha_i^- \cdot \big(-\lambda\pi_i^0(x)\big)^{\beta_i^-} + \gamma_i^- = -\alpha_i^- \cdot \big(-\pi_i^0(x)\big)^{\beta_i^-} \cdot \lambda^{\beta_i^-} + \gamma_i^-,$$

which is compatible with Eq. (15) only if $\beta_i^+ = \beta_i^- = \beta$ and $\gamma_i^+ = \gamma_i^- = \gamma_i$ for all i. Given the latter, we can again set $\gamma_i^+ = \gamma_i^- = 0$ by a positive affine transformation. As a result, the claim for both the logarithmic form and the power form is established for all contexts $\pi^0 \in \Pi^0$. □

Lemma 7 (Broad bracketing: Extension to contexts $\pi \notin \Pi^0$ and constant parameters)**.** *If $\succsim_\pi$ satisfies Axioms (1)–(4), (5) for all $\pi \in \Pi$, then there exist $\alpha_i^+, \alpha_i^- \neq 0$ ($i \leq n$) and $\beta \in \mathbb{R}$ such that for all $\pi \in \Pi$,*

$$v_{\pi,i}(p_i) = \beta \cdot \log p_i \qquad \textit{or} \qquad v_{\pi,i}(p_i) = \begin{cases} \alpha_i^+ \cdot (p_i)^\beta, & \textit{if } p_i \geq 0, \\ -\alpha_i^- \cdot (-p_i)^\beta, & \textit{if } p_i < 0. \end{cases}$$

Proof. By Axiom (5), broad bracketing, we know that preferences are scaling invariant in all contexts, fixing the functional form by Lemma 6, and it remains to show that $(\alpha_\pi^+, \alpha_\pi^-, \beta_\pi) = (\alpha_{\pi'}^+, \alpha_{\pi'}^-, \beta_{\pi'})$ for all $\pi, \pi' \in \Pi$. For any $\pi, \pi' \in \Pi$ with images such that their intersection $\pi[X] \cap \pi'[X]$ has positive mass in $\mathbb{R}^n$, equality of parameters follows immediately from the uniqueness of the solution to the functional equation solved in the proof of Lemma 6. Further, since the payoff images are cones in $\mathbb{R}^n$, $\pi[X] \cap \pi'[X]$ has positive mass in $\mathbb{R}^n$ if $\pi' = \pi + c$, for some $c \in \mathbb{R}^n$, if c is sufficiently close to $\mathbf{0}$. By the denseness of the rational numbers in $\mathbb{R}^n$, the claim follows by induction as in the proof of Lemma 4.

□

Lemma 8 (Narrow bracketing: Extension to contexts $\pi \notin \Pi^0$ and constant parameters)**.** *If $\succsim_\pi$ satisfies Axioms (1)–(3), (4), (6) and (7) for all $\pi \in \Pi$, then there exist $\alpha_i^+, \alpha_i^- \neq 0$ ($i \leq n$) and $\beta \in \mathbb{R}$ such that for all $\pi \in \Pi$,*

$$v_{\pi,i}(p_i) = \beta \cdot \log p_i \qquad \textit{or} \qquad v_{\pi,i}(p_i) = \begin{cases} \alpha_i^+ \cdot (p_i)^\beta, & \textit{if } p_i \geq 0, \\ -\alpha_i^- \cdot (-p_i)^\beta, & \textit{if } p_i < 0. \end{cases}$$

Proof. Fix any $\pi \in \Pi$, and fix the scaling invariant context $\pi^0 = S^0(\pi - \min \pi)$, which exists by Axiom (4). By Lemma 5, we know that

$$u_\pi(x) = \sum_{i \leq n} v_{\pi,i}\big(\pi_i(x) - r_i(\pi)\big)$$

represents $\succsim_\pi$ with $v_\pi = v_{\pi^0}$ as characterized in Lemma 6 where $r(\pi) = \min \pi - \min \pi^0$ and

$$v_{\pi,i}(p_i) = \beta_\pi \cdot \log p_i \qquad \text{or} \qquad v_{\pi,i}(p_i) = \begin{cases} \alpha_{\pi,i}^+ \cdot (p_i)^{\beta_\pi}, & \text{if } p_i \geq 0, \\ -\alpha_{\pi,i}^- \cdot (-p_i)^{\beta_\pi}, & \text{if } p_i < 0, \end{cases}$$

with $\alpha_{\pi,i}^+, \alpha_{\pi,i}^- \neq 0$ and $\beta_\pi \in \mathbb{R}$ for all $i \leq n$. Consequently, for all $\pi' \in \Pi$,

$$(\alpha_\pi^+, \alpha_\pi^-, \beta_\pi) = (\alpha_{\pi'}^+, \alpha_{\pi'}^-, \beta_{\pi'}) \quad \text{if} \quad \pi(0) - \min \pi = \pi'(0) - \min \pi'.$$

Now, given that

$$u_\pi(x) = \sum_{i \leq n} \alpha_i \cdot v_{\pi,i}\big[\pi_i(x) - r_i(\pi)\big],$$

represents $\succsim_\pi$, fix any $\pi' \in \Pi$ and note that by Compensability there exists $c \in \mathbb{R}^n$ such that $\pi' \sim \pi + c$. It follows that

$$u_{\pi'}(x') = \sum_{i \leq n} \alpha_i \cdot v_{\pi,i}\big[\pi'_i(x') - r_i(\pi')\big],$$

represents $\succsim'_\pi$ for set of points $x' \in X'$ satisfying $\pi'[X'] = \pi[X] \cap \pi'[X]$. Compensability implies that $\pi'[X']$ is non-empty, but since $\pi[X]$ and $\pi'[X]$ are cones in $\mathbb{R}^n$, there exists a translation of $\pi'' = \pi' + c'$, with c' in the neighborhood of $\mathbf{0}$, such that $\pi[X] \cap \pi''[X]$ has positive mass in $\mathbb{R}^n$. By Compensability, there exists c'' such that $\pi'' \sim \pi + c''$, and due to $\pi[X] \cap \pi''[X]$ having positive mass and the existence of a continuum of such translations of π' to entirely cover $\pi[X]$ with such intersections, and all of which have are connected to π' by narrow bracketing, this implies that the (unique) solutions to the functional equations defining the utility parameters are the same,

$$(\alpha^+_\pi, \alpha^-_\pi, \beta_\pi) = (\alpha^+_{\pi''}, \alpha^-_{\pi''}, \beta_{\pi''}) = (\alpha^+_{\pi'}, \alpha^-_{\pi'}, \beta_{\pi'})$$

for all $\pi, \pi' \in \Pi$ and any translation π'' of π', where $r(\pi') = r(\pi) + c = r(\pi + c)$ if $\pi' \sim \pi + c$. Hence, there exist $(\alpha^+, \alpha^-, \beta)$ such that $\alpha^+_i = \alpha^+_{\pi,i}$, $\alpha^-_i = \alpha^-_{\pi,i}$, and $\beta = \beta_\pi$ for all $i \leq n$ and all π. □

Lemma 9 (Characterizing reference points)**.** *If $\succsim_\pi$ satisfies Axioms (1)–(3), (4), (6) and (7)–(8) for all $\pi \in \Pi$, then there exists $w_{i,j} \in [0,1]$ such that for all $\pi \in \Pi$ and all $i \leq n$,*

$$r_i(\pi) = \min \pi_i + \sum_{j \neq n} w_{i,j} \cdot (\pi_j(0) - \min \pi_j).$$

Proof. Recall that for all $\pi, \pi' \in \Pi$ there exists $c \in \mathbb{R}^n$ such that $\pi' \sim \pi + c$, by Compensability, and that consequently, $r(\pi') = r(\pi + c)$ as well as $(\alpha, \beta, \lambda)_\pi = (\alpha, \beta, \lambda)_{\pi'}$. Hence, $r(\pi') = r(\pi) + c$, or $c = r(\pi') - r(\pi)$.

Next, pick any $\pi' \in \Pi$ such that $\pi' = \lambda\pi$ for some $\lambda \in \Lambda : \lambda \neq 1$ (which exists by context richness). By Compensability, there exists $c_\lambda \in \mathbb{R}^n$ such that $\pi' \sim \pi + c_\lambda$. Since

$$u_\pi(x') = \sum_{i \leq n} \alpha_i \cdot v_i\big[\pi_i(x') - r_i(\pi)\big],$$

represents $\succsim_\pi$, it follows by $\pi' = \lambda\pi$ that

$$u_{\pi'}(x') = \sum_{i \leq n} \alpha_i \cdot v_i\big[\lambda\pi_i(x') - r_i(\lambda\pi)\big],$$

represents $\succsim'_\pi$, and it follows by Compensability that

$$\tilde{u}_{\pi'}(x') = \sum_{i \leq n} \alpha_i \cdot v_i\big[\pi_i(x') + c_{\lambda,i} - r_i(\pi + c_\lambda)\big] = \sum_{i \leq n} \alpha_i \cdot v_i\big[\pi_i(x') - r_i(\pi)\big],$$

also represents $\succsim'_\pi$ for all $x' : \pi(x') \in \pi'[X] \cap \pi[X]$. This extends to the entire domain by an argument equivalent to that used in the proof of Lemma 8. Hence, $u_{\pi'}(x')$ and $\tilde{u}_{\pi'}(x')$ must be affine transformations of another, and given the functional form of (v_i), this implies $r_i(\lambda\pi) = \lambda r_i(\pi)$ for all $i \leq n$, $\pi \in \Pi$, and $\lambda \in \Lambda$.

Next, recall that by additive compensability,

$$\pi' \sim \pi + c' \text{ and } \pi'' \sim \pi + c'' \quad \Longrightarrow \quad \pi' + \pi'' \sim 2\pi + c' + c''.$$

Given $r(\pi + c) = r(\pi) + c$ and $r_i(\lambda\pi) = \lambda r_i(\pi)$, we obtain

$$r(\pi') = r(\pi) + c' \text{ and } r(\pi'') = r(\pi) + c'' \quad \Longrightarrow \quad r(\pi' + \pi'') = r(2\pi) + c' + c'' = 2r(\pi) + c' + c''$$

which implies, given that $2r(\pi)+c'+c''=\pi'+\pi''$ by assumption,

$$r(\pi'+\pi'')=r(\pi')+r(\pi''). \tag{20}$$

Now recall that by Lemma 5,

$$r(\pi)=\min\pi+f(\pi(0)-\min\pi)$$

for some function $f:\mathbb{R}^n\to\mathbb{R}^n$. Thus, $r(\pi'+\pi'')=r(\pi')+r(\pi'')$ implies $f(p+q)=f(p)+f(q)$ for all $p,q\in\{\pi(0)-\min\pi|\pi\in\Pi\}=:P$, since $p,q\in P$ implies $p+q\in P$ by context richness. By option richness and default richness, there exists $\overline{p}_1>0$ such that $(p_1,0,\ldots,0)\in P$ for all $p_1\leq\overline{p}_1$. Thus, $f(p+q)=f(p)+f(q)$ implies

$$f(p_1+q_1,0,\ldots,0)=f(p_1,0,\ldots,0)+f(q_1,0,\ldots,0) \tag{21}$$

for all $p_1,q_1,p_1+q_1\leq\overline{p}_1$. This is a Cauchy equation for $\tilde{f}(p_1)=f(p_1,0,\ldots,0)$ with the general solution

$$f(p_1,0,\ldots,0)=c_1p_1, \tag{22}$$

since $f(p_1,0,\ldots,0)$ is continuous by context continuity, and defined on an interval with positive length containing the element 0 (see Theorem 1 on page 46 of Aczél, 1966). Similarly, we obtain $f(0,\ldots,0,p_i,0,\ldots,0)=c_ip_i$ for all $i\leq n$, and by $r(\pi+\pi')=r(\pi)+r(\pi')$,

$$f(p_1,q_2,0,\ldots,0)=f(p_1,0,\ldots,0)+f(0,q_2,0,\ldots,0)=c_1\,p_1+c_2\,q_2. \tag{23}$$

By induction we extend this to all dimensions $i\leq n$ (see also Theorem 1 on page 215 of Aczél, 1966). We obtain (reverting to the original notation)

$$f_i(\pi(0)-\min\pi)=\sum_{j\leq n}c_j\cdot\big(\pi_j(0)-\min\pi_j\big),$$

and given $r(\pi)=\min\pi+f(\pi(0)-\min\pi)$ by Lemma 5 this implies (changing notation towards $w_{i,j}$),

$$r_i(\pi)=\min\pi_i+\sum_{j\leq n}w_{i,j}\cdot\big(\pi_j(0)-\min\pi_j\big),$$

for all $i\leq n$. □

□

B.2 Proofs of Propositions 2 and 3

B.2.1 Optimal choice of a regular dictator Δ in a given game Γ with $P_1=[0,B]$

Note that since for this part of the proof the game Γ is kept fixed, we drop the game index on the utility function and write r_i instead of $r_i(\Gamma)$ for the reference points. Then dictator Δ's utility function in game Γ is given by

$$u(p_1)=\frac{1}{\beta}\times\begin{Bmatrix}(p_1-r_1)^\beta & \text{if } p_1\geq r_1\\ -\delta(r_1-p_1)^\beta & \text{if } p_1<r_1\end{Bmatrix}+\frac{\alpha}{\beta}\times\begin{Bmatrix}(p_2(p_1)-r_2)^\beta & \text{if } p_2(p_1)\geq r_2\\ -\delta(r_2-p_2(p_1))^\beta & \text{if } p_2(p_1)<r_2\end{Bmatrix}$$

where $p_2(p_1)=t(B-p_1)$.

Step 1 Dictator Δ never chooses p_1 such that $p_1 < r_1$ and $p_2(p_1) < r_2$.

By satisfiability of reference points and $P_1 = [0,B]$ dictator Δ can always choose $p_1 \in P_1$ such that $p_1 \geq r_1$ and $p_2(p_1) \geq r_2$. This yields utility $u(p_1) = (p_1 - r_1)^\beta/\beta + \alpha(t(B-p_1) - r_2)^\beta/\beta \geq 0$ where the inequality follows by weak efficiency concerns ($0 < \beta < 1$). Choosing $p_1' \in P_1$ such that $p_1' < r_1$ and $p_2(p_1') < r_2$ instead yields utility $u(p_1') = -\delta(r_1 - p_1)^\beta/\beta - \alpha\delta(r_2 - t(B-p_1))^\beta/\beta < 0$.

Thus, we can restrict attention to the regions where at most one of the two players is in the loss-domain, i.e. does not reach her reference point. In the following we will first determine the local optima for dictator Δ in each of the three remaining regions. Then we can determine the global optimum by comparing utilities of the local optima.

Step 2 Local optimum in region 1: $p_1 \in [r_1, B - \frac{1}{t}r_2]$ ($\Leftrightarrow p_1 \geq r_1$ and $p_2(p_1) \geq r_2$)

The utility function that applies is

$$u^{(1)}(p_1) = (p_1 - r_1)^\beta + \alpha \cdot (t(B-p_1) - r_2)^\beta$$

Differentiating $u^{(1)}$ with respect to p_1 we get

$$\frac{du^{(1)}}{dp_1} = \beta\,(p_1 - r_1)^{\beta-1} - \alpha\beta t\,(t\,(B-p_1) - r_2)^{\beta-1}$$

which yields the first order condition

$$(p_1 - r_1)^{1-\beta}\,(t\,(B-p_1) - r_2)^{\beta-1} = \frac{1}{\alpha t} \qquad \Leftrightarrow \qquad \frac{t\,(B-p_1) - r_2}{p_1 - r_1} = (\alpha t)^{\frac{1}{1-\beta}}.$$

and the solution

$$p_1^+(\Gamma) = \frac{B + c_\alpha\, r_1 - r_2/t}{c_\alpha + 1} \quad \text{and} \quad p_2^+(\Gamma) = \frac{t c_\alpha\,(B - r_1) + r_2}{c_\alpha + 1}$$

using $c_\alpha := (\alpha t^\beta)^{\frac{1}{1-\beta}}$. Note that for $p_1 = B - \frac{1}{t}r_2$ and $p_1 = r_1$ the above first order condition is not defined because the utility function exhibits kinks at these points. We have $p_1^+(\Gamma) = B - \frac{1}{t}r_2 = r_1$ iff satisfiability is binding, i.e. $B - r_1 - \frac{1}{t}r_2 = 0$. By satisfiability we have $p_1^+(\Gamma) \in [r_1, B - \frac{1}{t}r_2]$ for all regular dictators Δ. Furthermore, the second order condition for $p_1^+(\Gamma)$ to be a maximum reduces to

$$t^{2-\beta} c_\alpha (1 + c_\alpha)^{1-\beta} (t(B - r_1) - r_2)^\beta > 0,$$

which is fulfilled for $p_1^+(\Gamma)$ by satisfiability, weak efficiency concerns ($0 < \beta < 1$), and $\alpha, t > 0$. Overall, we thus have for the local optimum in region 1

$$p_1^{(*)} = p_1^+(\Gamma).$$

Step 3 Local optimum in region 2: $p_1 \in (B - \frac{1}{t}r_2, B]$ ($\Leftrightarrow p_1 \geq r_1$ and $p_2 < r_2$)

The utility function that applies is

$$u^{(2)}(p_1) = (p_1 - r_1)^\beta - \delta\alpha \cdot (r_2 - t(B-p_1))^\beta$$

Differentiating $u^{(2)}$ with respect to p_1 we obtain

$$\frac{du^{(2)}}{dp_1} = \beta(p_1 - r_1)^{\beta-1} - \delta\alpha\beta t\,(r_2 - t\,(B-p_1))^{\beta-1}$$

which yields the first order condition

$$(p_1 - r_1)^{1-\beta}(r_2 - t(B-p_1))^{\beta-1} = \frac{1}{\delta\alpha t} \qquad \Leftrightarrow \qquad \frac{r_2 - t(B-p_1)}{p_1 - r_1} = (\delta\alpha t)^{\frac{1}{1-\beta}}$$

and the solution

$$p_1^{(2)}(\Gamma) = \frac{B - \delta^{\frac{1}{1-\beta}} c_\alpha r_1 - r_2/t}{1 - \delta^{\frac{1}{1-\beta}} c_\alpha} \quad \text{and} \quad p_2^{(2)}(\Gamma) = \frac{t\delta^{\frac{1}{1-\beta}} c_\alpha (r_1 - B) + r_2}{1 - \delta^{\frac{1}{1-\beta}} c_\alpha}.$$

By satisfiability we have $p_1^{(2)} \in (B - \frac{1}{t} r_2, B]$ iff $\delta^{\frac{1}{1-\beta}} c_\alpha \leq \frac{r_2}{t(B-r_1)} \Leftrightarrow \delta \leq \frac{1}{\alpha t^\beta} \left(\frac{r_2}{t(B-r_1)}\right)^{1-\beta}$. Using $\delta^{\frac{1}{1-\beta}} c_\alpha < 1$, the second order condition for $p_1^{(2)}(\Gamma)$ to be a maximum reduces to

$$t^{2-\beta} \delta^{\frac{1}{1-\beta}} c_\alpha (1 - \delta^{\frac{1}{1-\beta}} c_\alpha)^{1-\beta} (t(B-r_1) - r_2)^\beta < 0.$$

Thus, the second order condition does not hold for any $p_1^{(2)}(\Gamma) \in (B - \frac{1}{t} r_2, B]$ by satisfiability and weak efficiency concerns ($0 < \beta < 1$). It follows that the local optimum is either $p_1 = B - \frac{1}{t} r_2$ or $p_1 = B$ depending on whether $u^{(2)}(B - \frac{1}{t} r_2) \geq u^{(2)}(B)$, a condition which reduces to

$$\delta \geq c_\alpha^{\beta-1} \left(\left(\frac{t(B-r_1)}{r_2} \right)^\beta - \left(\frac{t(B-r_1)}{r_2} - 1 \right)^\beta \right).$$

Overall, we thus have for the local optimum in region 2

$$p_1^{(*)} = \begin{cases} B - \frac{1}{t} r_2 & \text{if } \delta \geq c_\alpha^{\beta-1} \left(\left(\frac{t(B-r_1)}{r_2} \right)^\beta - \left(\frac{t(B-r_1)}{r_2} - 1 \right)^\beta \right) \\ B & \text{else.} \end{cases}$$

Step 4 Local optimum in region 3: $p_1 \in [0, r_1)$ ($\Leftrightarrow p_1 < r_1$ and $p_2(p_1) \geq r_2$)

The utility function that applies is

$$u^{(3)}(p_1) = -\delta \cdot (r_1 - p_1)^\beta + \alpha \cdot (t(B-p_1) - r_2)^\beta$$

Differentiating $u^{(3)}$ with respect to p_1 we obtain

$$\frac{du^{(3)}}{dp_1} = \delta\beta(r_1 - p_1)^{\beta-1} - \alpha\beta t(t(B-p_1) - r_2)^{\beta-1}$$

which yields the first order condition

$$(r_1 - p_1)^{1-\beta}(t(B-p_1) - r_2)^{\beta-1} = \frac{\delta}{\alpha t} \qquad \Leftrightarrow \qquad \frac{t(B-p_1) - r_2}{r_1 - p_1} = \left(\frac{\alpha t}{\delta}\right)^{\frac{1}{1-\beta}}$$

and the solution

$$p_1^{(3)}(\Gamma) = \frac{B - \delta^{1-\beta} c_\alpha r_1 - r_2/t}{1 - \delta^{1-\beta} c_\alpha} \quad \text{and} \quad p_2^{(3)}(\Gamma) = \frac{t\delta^{1-\beta} c_\alpha (r_1 - B) + r_2}{1 - \delta^{1-\beta} c_\alpha}$$

By satisfiability we have $p_1^{(3)} \in [0, r_1)$ iff $\delta^{1-\beta} c_\alpha \geq \frac{tB - r_2}{t r_1} \Leftrightarrow \delta \leq \alpha t^\beta \left(\frac{t r_1}{tB - r_2}\right)^{1-\beta}$. The second order

condition for $p_1^{(3)}(\Gamma)$ to be a maximum reduces to

$$\frac{1}{\delta^{1-\beta}c_\alpha}\left(\frac{r_1-B+r_2/t}{1-\delta^{1-\beta}c_\alpha}\right)^{\beta-2} > \left(\frac{r_1-B+r_2/t}{1-\delta^{1-\beta}c_\alpha}\right)^{\beta-2},$$

which by satisfiability does not hold for any $p_1^{(3)}(\Gamma) \in [0, r_1)$. It follows that the optimum is either $p_1 = 0$ or $p_1 = r_1$ depending on whether $u^{(3)}(0) \geq u^{(3)}(r_1)$, a condition which reduces to

$$\delta \leq c_\alpha^{1-\beta}\left(\left(\frac{tB-r_2}{tr_1}\right)^\beta - \left(\frac{tB-r_2}{tr_1}-1\right)^\beta\right).$$

Overall, we thus have for the local optimum in region 3

$$p_1^{(*)} = \begin{cases} 0 & \text{if } \delta \leq c_\alpha^{1-\beta}\left(\left(\frac{tB-r_2}{tr_1}\right)^\beta - \left(\frac{tB-r_2}{tr_1}-1\right)^\beta\right) \\ r_1 & \text{else.} \end{cases}$$

Step 5 Reducing the set of candidate solutions for the global optimum

Using weak efficiency concerns ($0 < \beta < 1$) and $\alpha, t \geq 0$ we have $u(p_1^+(\Gamma)) \geq u(B - \frac{1}{t}r_2)$ and $u(p_1^+(\Gamma)) \geq u(r_1)$ for all regular dictators Δ, a result which obtains by simple rearrangement of the two inequalities. Thus, the remaining candidate solutions for the overall utility maximizer are $p_1 = p_1^+(\Gamma)$, $p_1 = B$, and $p_1 = 0$.

Furthermore, we have $u(p_1^+(\Gamma)) \geq u(0)$ iff

$$\delta \geq c_\alpha^{1-\beta}\left(\frac{tB-r_2}{tr_1}\right)^\beta - (c_\alpha+1)^{1-\beta}\left(\frac{tB-r_2}{tr_1}-1\right)^\beta. \tag{24}$$

From weak efficiency concerns ($0 < \beta < 1$) we can conclude that

$$c_\alpha^{1-\beta} < (c_\alpha+1)^{1-\beta}.$$

Define $f(x) = x^\beta$, then weak efficiency concerns ($0 < \beta < 1$) imply that f is subadditive in the domain $\mathbb{R}^+$, i.e. $f(a)+f(b) \geq f(a+b) \forall a, b \geq 0$. Thus, using satisfiability and letting $a = \frac{tB-r_2}{tr_1} - 1$ and $b = 1$, we have

$$f(a)+f(b) = \left(\frac{tB-r_2}{tr_1}-1\right)^\beta + 1^\beta \geq \left(\frac{tB-r_2}{tr_1}\right)^\beta = f(a+b)$$

implying

$$\left(\frac{tB-r_2}{tr_1}\right)^\beta - \left(\frac{tB-r_2}{tr_1}-1\right)^\beta \leq 1.$$

Suppose $c_\alpha^{1-\beta} \leq 1$. In this case we can conclude that the lower bound for δ defined in (24) is lower or equal 1, which by weak loss aversion ($\delta \geq 1$) implies $u(p_1^+(\Gamma)) \geq u(0)$. Note that $c_\alpha^{1-\beta} \leq 1$ by weak altruism ($0 \leq \alpha \leq 1$) always holds under no efficiency gains from giving ($t \leq 1$) such that in this case the candidate solutions for the overall utility maximizer reduce further to $p_1 = p_1^+(\Gamma)$ and $p_1 = B$.

Finally, we have $u(p_1^+(\Gamma)) \geq u(B)$ iff

$$\delta \geq c_\alpha^{\beta-1}\left(\left(\frac{t(B-r_1)}{r_2}\right)^\beta - (c_\alpha+1)^{1-\beta}\left(\frac{t(B-r_1)}{r_2}-1\right)^\beta\right). \tag{25}$$

Suppose $c_\alpha^{1-\beta} > 1$. In this case by a similar argument as above we can conclude that the lower bound for δ defined in (25) is lower or equal 1, which by weak loss aversion ($\delta \geq 1$) implies $u(p_1^+(\Gamma)) \geq u(B)$. We can therefore conclude that under efficiency gains from giving ($t > 1$) the candidate solutions for the overall utility maximizer reduce to $p_1 = p_1^+(\Gamma)$ and $p_1 = B$ in case $c_\alpha^{1-\beta} \leq 1$ while they reduce to $p_1 = p_1^+(\Gamma)$ and $p_1 = 0$ in case $c_\alpha^{1-\beta} > 1$.

Step 6 Global optimum

For the global optimum we have to distinguish the following two cases:

- Case 1: $c_\alpha^{1-\beta} \leq 1$

$$p_1^* = \begin{cases} p_1^+(\Gamma) & \text{if } \delta \geq \delta^+(\Gamma) \\ B & \text{else.} \end{cases}$$

with

$$\delta^+(\Gamma) := c_\alpha^{\beta-1} \left(\left(\frac{t(B-r_1)}{r_2} \right)^\beta - (c_\alpha + 1)^{1-\beta} \left(\frac{t(B-r_1)}{r_2} - 1 \right)^\beta \right)$$

- Case 2: $c_\alpha^{1-\beta} > 1$

$$p_1^* = \begin{cases} p_1^+(\Gamma) & \text{if } \delta \geq \delta^-(\Gamma) \\ 0 & \text{else.} \end{cases}$$

with

$$\delta^-(\Gamma) := c_\alpha^{1-\beta} \left(\frac{tB-r_2}{tr_1} \right)^\beta - (c_\alpha + 1)^{1-\beta} \left(\frac{tB-r_2}{tr_1} - 1 \right)^\beta$$

Note that under no efficiency gains from giving ($t \leq 1$) only case 1 applies.

B.2.2 Establishing the comparative statics

Step 1 **Non-convexity** In any game Γ with $P_1 = [0,B]$ there are dictators with non-convex preferences.

Fix a game Γ with $P_1 = [0,B]$. Consider a dictator Δ with $\delta \leq \bar{\delta}(\Gamma)$ where

$$\bar{\delta}(\Gamma) := c_\alpha^{\beta-1} \left(\frac{r_2(\Gamma)}{t(B-r_1(\Gamma))} \right)^{1-\beta}.$$

We have shown in step 3 of A.2.1 that the utility function of this dictator attains a minimum at $p_1 = p_1^{(2)}(\Gamma) \in [B - r_2/t, B]$ and has no other local extrema in that region. Furthermore, we have shown in step 2 of A.2.1 that her utility function attains a maximum at $p_1 = p_1^+(\Gamma) \in [r_1, B - r_2/t]$ and has no other local extrema in that region. Consider options a and b with $p_1^a = B$ and $p_1^b = p_1^+(\Gamma)$. Construct option c by choosing $\lambda \in [0,1]$ such that $p_1^c = \lambda p_1^a + (1-\lambda) p_1^b = p_1^{(2)}(\Gamma)$. Then, for dictator Δ in game Γ there exists an option d with $p_1^d \in (p_1^+(\Gamma), B)$ such that $u_\Gamma(p_1^a) \geq u_\Gamma(p_1^d)$ and $u_\Gamma(p_1^b) \geq u_\Gamma(p_1^d)$ but $u_\Gamma(p_1^c) < u_\Gamma(p_1^d)$. Since u_Γ represents dictator Δ's preferences in game Γ, this implies that her preferences are non-convex.

We still have to show that in any game Γ with $P_1 = [0,B]$ there exist regular dictators with $\delta \leq \bar{\delta}(\Gamma)$. For any transfer rate t specified by Γ we can find (α, β) satisfying $0 \leq \alpha \leq 1$ and $0 < \beta < 1$ such that $c_\alpha^{1-\beta} \leq 1 \Leftrightarrow c_\alpha^{\beta-1} \geq 1$. Given such (α, β), for any endowments (B_1, B_2) specified by Γ, we can find (w_1, w_2) in accordance with satisfiability resulting in reference points

$r_1(\Gamma) = w_1B_1 + w_2B_2$ and $r_2(\Gamma) = t(w_1B_2 + w_2B_1)$ such that $r_2(\Gamma)/t(B - r_1(\Gamma))$ is close enough to 1 to make $\bar{\delta}(\Gamma) \geq 1$. Thus, given such $(\alpha, \beta, w_1, w_2)$, we can conclude that there exist δ satisfying weak loss aversion ($\delta \geq 1$) such that $\delta \leq \bar{\delta}(\Gamma)$.

Step 2 Taking options reduce giving both at the extensive and intensive margin Introducing a taking option turns some initial givers into takers and reduces average amounts given.

Consider two games $\Gamma = \langle B_1, B_2, P_1, t\rangle$ and $\Gamma' = \langle B_1, B_2, P_1', t\rangle$ with $B_2 > 0$ that are equivalent in every dimension except the choice set of the dictator. In Γ the choice set is restricted to $P_1 = [0, \max p_1]$ with $\max p_1 = B_1$ and in Γ' the choice set is extended to $p_1' = [0, \max p_1']$ with $B_1 < \max p_1' \leq B_1 + B_2$.

Moving from Γ to Γ' the only game parameter that changes is the maximum payoff for the dictator which rises from $\max p_1 = B_1$ to $\max p_1'$. As a result of this rise, the minimum payoff for the recipient adjusts accordingly, i.e. it falls from $\min p_2 = t(B_1 + B_2 - \max p_1) = tB_2$ to $\min p_2' = t(B_1 + B_2 - \max p_1')$. Therefore, the utility functions of a regular dictator Δ in Γ and Γ' differ in the players' reference points. We have

$$r_2(\Gamma) = t\left(B_2 + (1 - w_1 + w_2)B_1 - (1 - w_1)\max p_1\right) \quad \text{with} \quad \frac{dr_2}{d\max p_1} = -t(1 - w_1) \leq 0,$$

and

$$r_1(\Gamma) = (w_1 - w_2)B_1 + w_2 \max p_1 \quad \text{with} \quad \frac{dr_1}{d\max p_1} = w_2 \geq 0,$$

where the inequalities follow from satisfiability. Thus, we have $r_2(\Gamma) \geq r_2(\Gamma')$ and $r_1(\Gamma) \leq r_1(\Gamma')$. Plugging in our reference points we get for the interior solution in game Γ

$$p_1^+(\Gamma) = (w_1 - w_2)B_1 + \frac{1 - w_1 + c_\alpha w_2}{c_\alpha + 1}\max p_1$$

and the derivative with respect to the maximum payoff of the dictator is given by

$$\frac{dp_1^+}{d\max p_1} = \frac{1 - w_1 + c_\alpha w_2}{c_\alpha + 1} \geq 0$$

where the inequality follows from $\alpha \geq 0$ and satisfiability. Thus, we have $p_1^+(\Gamma) \leq p_1^+(\Gamma')$. Note furthermore, that by satisfiability $\frac{dp_1^+}{d\max p_1} \leq 1$ implying that the interior solution is feasible for any regular dictator in Γ and Γ'.

In A.2.1. we specified the global optimum for games like Γ with $P_1 = [0, B]$. In games like Γ' where the choice set of the dictator is restricted to $P_1' = [0, \max p_1]$ with $\max p_1 < B$ the selfish corner solution $p_1 = B$ is not feasible. Thus, we have for $c_\alpha^{1-\beta} \leq 1$ (case 1):

$$p_1^* = \begin{cases} p_1^+(\Gamma) & \text{if } \delta \geq \hat{\delta}^+(\Gamma) \\ \max p_1 & \text{else.} \end{cases}$$

with

$$\hat{\delta}^+(\Gamma) := c_\alpha^{\beta-1}\left(\left(\frac{t(\max p_1 - r_1)}{r_2 - t(B - \max p_1)}\right)^\beta - (c_\alpha + 1)^{1-\beta}\left(\frac{t(B - r_1) - r_2}{r_2 - t(B - \max p_1)}\right)^\beta\right)$$

where the expression for $\hat{\delta}^+(\Gamma)$ follows from rearrangement of $u_\Gamma(p_1^+(\Gamma)) \geq u_\Gamma(\max p_1)$. Note that for $c_\alpha^{1-\beta} > 1$ (case 2) the specification of the global optimum is not affected by the restriction of the choice set because the altruistic corner solution $p_1 = 0$ is feasible in Γ'.

We consider this threshold $\hat{\delta}^+(\Gamma')$ such that in game Γ' among the regular dictators with

$c_\alpha^{1-\beta} \leq 1$, those with $\delta < \hat{\delta}^+(\Gamma')$ choose the selfish corner solution $p_1 = \max p_1'$ while those with $\delta \geq \hat{\delta}^+(\Gamma')$ choose the interior solution $p_1 = p_1^+(\Gamma')$. We can rewrite it as

$$\hat{\delta}^+(\Gamma') := c_\alpha^{\beta-1}\left(\left(\frac{(1-w_2)\max p_1' - (w_1-w_2)B_1}{w_1 \max p_1' - (w_1-w_2)B_1}\right)^\beta - (c_\alpha+1)^{1-\beta}\left(\frac{(1-w_2)\max p_1' - (w_1-w_2)B_1}{w_1 \max p_1' - (w_1-w_2)B_1} - 1\right)^\beta\right).$$

Then the derivative with respect to $\max p_1'$ is given by

$$\frac{d\hat{\delta}^+}{d\max p_1'} = \frac{\beta(1-w_1-w_2)(w_1-w_2)B_1}{c_\alpha^{1-\beta}(w_1 \max p_1' - (w_1-w_2)B_1)^2}\left((c_\alpha+1)^{1-\beta}\left(\frac{(1-w_2)\max p_1' - (w_1-w_2)B_1}{w_1 \max p_1' - (w_1-w_2)B_1} - 1\right)^{\beta-1} - \left(\frac{(1-w_2)\max p_1' - (w_1-w_2)B_1}{w_1 \max p_1' - (w_1-w_2)B_1}\right)^{\beta-1}\right).$$

From weak altruism, weak efficiency concerns, satisfiability, and $w_1 \geq w_2$ we can conclude that $\frac{d\hat{\delta}^+}{d\max p_1'} \geq 0$. Thus, we have $\hat{\delta}^+(\Gamma) \leq \hat{\delta}^+(\Gamma')$, implying that weakly more regular dictators with $c_\alpha^{1-\beta} \leq 1$ prefer the selfish corner solution to the interior solution in Γ' compared to Γ.

Now, consider the threshold $\delta^-(\Gamma)$ such that in game Γ among the regular dictators with $c_\alpha^{1-\beta} > 1$, those with $\delta < \delta^-(\Gamma)$ prefer the altruistic corner solution $p_1 = 0$ while those with $\delta \geq \delta^-(\Gamma)$ prefer the interior solution $p_1 = p_1^+(\Gamma)$. We can rewrite the threshold as

$$\delta^-(\Gamma) = c_\alpha^{1-\beta}\left(\frac{(1-w_1)\max p_1 + (w_1-w_2)B_1}{w_2 \max p_1 + (w_1-w_2)B_1}\right)^\beta - (c_\alpha+1)^{1-\beta}\left(\frac{(1-w_1)\max p_1 + (w_1-w_2)B_1}{w_2 \max p_1 + (w_1-w_2)B_1} - 1\right)^\beta.$$

Then the derivative with respect to $\max p_1$ is given by

$$\frac{d\delta^-}{d\max p_1} = \frac{\beta(1-w_1-w_2)(w_1-w_2)B_1}{(w_2 \max p_1 + (w_1-w_2)B_1)^2}\left(c_\alpha^{1-\beta}\left(\frac{(1-w_1)\max p_1 + (w_1-w_2)B_1}{w_2 \max p_1 + (w_1-w_2)B_1}\right)^{\beta-1} - (c_\alpha+1)^{1-\beta}\left(\frac{(1-w_1)\max p_1 + (w_1-w_2)B_1}{w_2 \max p_1 + (w_1-w_2)B_1} - 1\right)^{\beta-1}\right).$$

From weak altruism, weak efficiency concerns, satisfiability, and $w_1 \geq w_2$ we can conclude that $\frac{d\delta^-}{d\max p_1} \leq 0$. Thus, we have $\delta^-(\Gamma) \geq \delta^-(\Gamma')$ implying that weakly less regular dictators with $c_\alpha^{1-\beta} > 1$ prefer the altruistic corner solution to the interior solution in Γ' compared to Γ.

Using these results together with our results from A.2.1 we can show that comparing the choice of any regular dictator Δ in Γ to her choice in Γ' one of the following cases applies:

(i) Her choice switches from $p_1 = p_1^+(\Gamma)$ to $p_1 = p_1^+(\Gamma')$ where $p_1^+(\Gamma) \leq p_1^+(\Gamma')$.

(ii) Her choice switches from $p_1 = p_1^+(\Gamma)$ to $p_1 = \max p_1'$ where $p_1^+(\Gamma) < \max p_1'$.

(iii) Her choice switches from $p_1 = 0$ to $p_1 = p_1^+(\Gamma')$ where $0 \leq p_1^+(\Gamma')$.

(iv) Her choice remains at $p_1 = 0$.

First, we restrict attention to regular dictators with $c_\alpha^{1-\beta} \leq 1$. Note that in game Γ by satisfiability $r_2(\Gamma) \leq B_2$ such that there is no feasible choice for the dictator in which the recipient's reference point is not fulfilled. Thus, in game Γ these dictators all choose the interior solution $p_1 = p_1^+(\Gamma)$. Now consider the same dictators in Γ' and split them into two groups according to their loss aversion parameters. The dictators with $\delta \geq \hat{\delta}^+(\Gamma')$ choose $p_1 = p_1^+(\Gamma')$ in Γ'. The dictators with $\delta < \hat{\delta}^+(\Gamma')$ choose $p_1 = \max p_1'$ in Γ'.

Now, restrict attention to regular dictators with $c_\alpha^{1-\beta} > 1$. We split these dictators into three groups according to their loss aversion parameters. Consider first the dictators with $\delta \geq \delta^-(\Gamma)$. These dictators choose $p_1 = p_1^+(\Gamma)$ in Γ. Since $\delta^-(\Gamma) \geq \delta^-(\Gamma')$ they choose $p_1 = p_1^+(\Gamma')$ in Γ'. Second, consider the dictators with $\delta \in [\delta^-(\Gamma'), \delta^-(\Gamma))$. These dictators choose $p_1 = 0$ in Γ and switch to $p_1 = p_1^+(\Gamma')$ in Γ'. Third, consider the dictators with $\delta < \delta^+(\Gamma')$. These dictators choose $p_1 = 0$ both in Γ and in Γ'.

We still have to show that for any Γ and Γ' there exist regular dictators who give in Γ and switch to taking in Γ'. We show that for any Γ and Γ' there exist regular dictators with $p_1^+(\Gamma) < B_1$ and $\delta < \hat{\delta}^+(\Gamma')$, i.e. regular dictators who give at the interior solution in Γ and to whom case (ii) applies. We have $p_1^+(\Gamma) < B_1$ iff $c_\alpha(1-w_1) + w_2 > 0$. Thus, we have $p_1^+(\Gamma) < B_1$ for all

regular dictators with $0 < w_1 < 1$ or $w_1 = 1$ and $w_2 > 0$. Now, for any transfer rate t specified by Γ and Γ' we can find (α, β) satisfying $0 < \alpha \leq 1$ and $0 < \beta < 1$ such that $c_\alpha^{1-\beta} < 1 \Leftrightarrow c_\alpha^{\beta-1} > 1$. Given such (α, β), for any endowments and choice set (B_1, B_2, P_1') specified by Γ' we have $((1-w_2)\max p_1' - (w_1 - w_2)B_1)/(w_1 \max p_1' - (w_1 - w_2)B_1) = 1$ for $w_1 = 1$ and $w_2 = 0$. Thus, by continuity of $\hat{\delta}^+$ we can always find $w_1 > 0$ and $w_2 \geq 0$ in accordance with satisfiability such that the expression is close enough to 1 to make $\hat{\delta}^+(\Gamma') > 1$ and given such $(\alpha, \beta, w_1, w_2)$, we can conclude that there exist δ satisfying weak loss aversion ($\delta \geq 1$) such that $\delta < \hat{\delta}^+(\Gamma')$.

Finally, we need to show that for any Γ and Γ' there exist regular dictators who give more in Γ than in Γ'. We show that for any Γ and Γ' there exist regular dictators with $p_1^+(\Gamma), p_1^+(\Gamma') < B_1$ and $\delta \geq \hat{\delta}^+(\Gamma')$. As above we have $p_1^+(\Gamma) < B_1$ for all regular dictators with $0 < w_1 < 1$ or $w_1 = 1$ and $w_2 > 0$. Furthermore, we have $\frac{dp_1^+}{d \max p_1} = 0$ for $w_1 = 1$ and $w_2 = 0$. Thus, by continuity of $p_1^+(\Gamma)$ for any transfer rate t specified by Γ and Γ' we can find $0 < w_1 \leq 1$ and $w_2 \geq 0$ in accordance with satisfiability such that $p_1^+(\Gamma) < p_1^+(\Gamma') < B_1$. Since there is no upper bound on the loss aversion parameter of regular dictators given such (w_1, w_2) there always exist regular dictators with $\delta \geq \hat{\delta}^+(\Gamma')$.

Step 3 Incomplete crowding out Reallocating initial endowment from dictator to recipient results (in expectation) in a payoff increase for the recipient.

Consider two games $\Gamma = \langle B_1, B_2, P_1, t \rangle$ and $\Gamma' = \langle B_1', B_2', P_1', t \rangle$ without taking option, i.e. $P_1 = [0, B_1]$ and $P_1' = [0, B_1']$, where Γ' is generated from Γ by reallocating initial endowment from the dictator to the recipient, i.e. $B_1 + B_2 = B_1' + B_2' = \bar{B}$ and $B_1 < B_1'$. Thus, comparing such games we can write the recipient's endowment as a function of the dictator's endowment, i.e. $B_2(B_1) = \bar{B} - B_1$.

Moving from Γ to Γ' the game parameters that change are the player's endowments and the maximum payoff for the dictator. The dictator's endowment falls from B_1 to B_1' while the recipient's endowment rises from $\bar{B} - B_1$ to $\bar{B} - B_1'$. Furthermore, the maximum payoff for the dictator falls from B_1 to B_1' such that the minimum payoff for the recipient rises from $\min p_2 = t(\bar{B} - B_1)$ to $\min p_2' = t(\bar{B} - B_1')$. Therefore, the utility functions of a regular dictator Δ in Γ and Γ' differ in the reference points of the dictator and the recipient. We have

$$r_1(\Gamma) = w_1 B_1 \quad \text{with} \quad \frac{dr_1}{dB_1} = w_1 \geq 0,$$

where the inequality follows from satisfiability, and

$$r_2(\Gamma) = t(\bar{B} - (1 - w_2)B_1) \quad \text{with} \quad \frac{dr_2}{dB_1} = -t(1 - w_2) \leq 0,$$

where the inequality follows from satisfiability and $t > 0$. Thus, we have $r_1(\Gamma) \geq r_1(\Gamma')$ and $r_2(\Gamma) \leq r_2(\Gamma')$. We can rewrite the interior solution as

$$p_1^+(\Gamma) = \frac{1 + c_\alpha w_1 - w_2}{c_\alpha + 1} B_1.$$

Taking the derivative with respect to the dictator's initial endowment we get

$$\frac{dp_1^+}{dB_1} = \frac{1 + c_\alpha w_1 - w_2}{c_\alpha + 1} \geq 0$$

where the inequality follows from imperfect altruism, weak efficiency concerns, and satisfiability. Thus, we have $p_1^+(\Gamma) \geq p_1^+(\Gamma')$.

Consider now the threshold for $\delta^-(\Gamma)$ such that in game Γ among the regular dictators with $c_\alpha^{1-\beta} > 1$, those with $\delta < \delta^-(\Gamma)$ choose the altruistic corner solution $p_1 = 0$ while those with

$\delta \geq \delta^-(\Gamma)$ choose the interior solution $p_1 = p_1^+(\Gamma)$. We can rewrite the threshold as

$$\delta^-(\Gamma) = c_\alpha^{1-\beta}\left(\frac{1-w_2}{w_1}\right)^\beta - (c_\alpha+1)^{1-\beta}\left(\frac{1-w_2}{w_1}-1\right)^\beta$$

and since the threshold is independent of B_1 we get $\frac{d\delta^-}{dB_1} = 0$. Thus, we have $\delta^-(\Gamma) = \delta^-(\Gamma') =: \delta^-$.

Using these results together with our results from A.2.1 we can show that comparing the choice of any regular dictator Δ in Γ to her choice in Γ' one of the following cases applies:

(i) Her choice switches from $p_1 = p_1^+(\Gamma)$ to $p_1 = p_1^+(\Gamma')$ where $p_1^+(\Gamma) \geq p_1^+(\Gamma')$.

(ii) Her choice remains at $p_1 = 0$.

Consider first only regular dictators with $c_\alpha^{1-\beta} \leq 1$. Since in neither Γ nor Γ' there is a feasible choice such that the reference point of the recipient is not fulfilled, these dictators all choose the respective interior solution in Γ and Γ'.

Now, consider regular dictators with $c_\alpha^{1-\beta} > 1$. We split these dictators into two groups according to their loss aversion parameters. Consider first the dictators with $\delta \geq \delta^-$. These dictators choose $p_1 = p_1^+(\Gamma)$ in Γ and $p_1 = p_1^+(\Gamma')$ in Γ'. Second, consider the dictators with $\delta < \delta^-$. These dictators choose $p_1 = 0$ both in Γ and Γ'.

Finally, we show that for any Γ and Γ' there exist regular dictators to whom case (i) applies in a strict sense, i.e. regular dictators whose choice in Γ' compared to Γ strictly increases the payoff of the recipient. For any transfer rate t specified by Γ and Γ' we can find $\alpha > 0$ and β satisfying weak altruism and weak efficiency concerns such that $c_\alpha^{1-\beta} \leq 1$, i.e. for any transfer rate t we can find regular dictators to whom case (i) applies. Furthermore, given such (α,β) we can always find (w_1,w_2) in accordance with satisfiability such that $dp_1^+/dB_1 > 0$.

Step 4 Efficiency concerns The recipient's payoff is weakly increasing in the transfer rate.

Consider two games $\Gamma = \langle B_1,B_2,P_1,t\rangle$ and $\Gamma' = \langle B_1,B_2,P_1,t'\rangle$ with $t < t'$, $P_1 = [0,\max p_1]$, and $B_1 \leq \max p_1 \leq B_1 + B_2$ which are equivalent in every dimension except the transfer rate.

The utility functions of a regular dictator Δ in Γ and Γ' differ only in the reference points of the recipient. His endowment is multiplied with t' instead of t and his minimal payoff increases from $\min p_2 = t(B - \max p_1)$ to $\min p_2' = t'(B - \max p_1)$. We have

$$r_2(\Gamma) = t(B_2 + (1-w_1+w_2)B_1 - (1-w_1)\max p_1)$$

with

$$\frac{dr_2}{dt} = B_2 + (1-w_1+w_2)B_1 - (1-w_1)\max p_1 \geq 0$$

where the inequality follows by satisfiability and $\max p_1 \leq B_1 + B_2$. Thus, we have $r_2(\Gamma) \leq r_2(\Gamma')$. We can rewrite the interior solution as

$$p_1^+(\Gamma) = \frac{t((1-w_1)\max p_1 + (w_1-w_2)B_1) + (\alpha t)^{\frac{1}{1-\beta}} r_1(\Gamma)}{(\alpha t)^{\frac{1}{1-\beta}} + t}.$$

Taking the derivative with respect to the transfer rate we get

$$\frac{dp_1^+}{dt} = \frac{tc_\alpha\beta}{1-\beta}(r_1(\Gamma) - (1-w_1)\max p_1 - (w_1-w_2)B_1) = \frac{tc_\alpha\beta}{1-\beta}(w_1+w_2-1)\max p_1 \leq 0$$

where the inequality follows from weak altruism, weak efficiency concerns, and satisfiability. Thus, we have $p_1^+(\Gamma) \geq p_1^+(\Gamma')$.

Consider now the threshold $\hat{\delta}^+(\Gamma)$ such that in a game Γ with $\max p_1 > B_1$ among the regular dictators with $c_\alpha^{1-\beta} \leq 1$, those with $\delta < \hat{\delta}^+(\Gamma)$ choose the selfish corner solution $p_1 = \max p_1$ while those with $\delta \geq \hat{\delta}^+(\Gamma)$ choose the interior solution $p_1 = p_1^+(\Gamma)$. We can rewrite this threshold as

$$\hat{\delta}^+(\Gamma) = \frac{1}{\alpha t^\beta}\left(\left(\frac{\max p_1 - r_1}{w_1 \max p_1 - (w_1 - w_2)B_1}\right)^\beta - \left(\left(\alpha t^\beta\right)^{\frac{1}{1-\beta}} + 1\right)^{1-\beta}\left(\frac{\max p_1 - r_1}{w_1 \max p_1 - (w_1 - w_2)B_1} - 1\right)^\beta\right).$$

Taking the derivative with respect to t we get

$$\frac{d\hat{\delta}^+}{dt} = \frac{\beta}{t c_\alpha^{1-\beta}}\left((c_\alpha + 1)^{-\beta}\left(\frac{\max p_1 - r_1}{w_1 \max p_1 - (w_1 - w_2)B_1} - 1\right)^\beta - \left(\frac{\max p_1 - r_1}{w_1 \max p_1 - (w_1 - w_2)B_1}\right)^\beta\right) \leq 0,$$

where the inequality follows from weak altruism, weak efficiency concerns, and satisfiability. Thus, we have $\hat{\delta}^+(\Gamma) \geq \hat{\delta}^+(\Gamma')$, implying that weakly more regular dictators with $c_\alpha^{1-\beta} \leq 1$ choose the selfish corner solution in Γ compared to Γ'.

Consider now the threshold $\delta^-(\Gamma)$ such that in game Γ among the regular dictators with $\alpha t^\beta > 1$, those with $\delta < \delta^-(\Gamma)$ choose the altruistic corner solution $p_1 = 0$ while those with $\delta \geq \delta^-(\Gamma)$ choose the interior solution $p_1 = p_1^+(\Gamma)$. We can rewrite this threshold as

$$\delta^-(\Gamma) = \alpha t^\beta\left(\frac{(1 - w_1)\max p_1 + (w_1 - w_2)B_1}{r_1}\right)^\beta - \left(\left(\alpha t^\beta\right)^{\frac{1}{1-\beta}} + 1\right)^{1-\beta}\left(\frac{(1 - w_1)\max p_1 + (w_1 - w_2)B_1}{r_1} - 1\right)^\beta.$$

Taking the derivative with respect to t we get

$$\frac{d\delta^-}{dt} = \frac{\beta}{t}\left(c_\alpha^{1-\beta}\left(\frac{B_1 - (1 - w_2)(\max p_1 - B_1)}{r_1}\right)^\beta - \frac{c_\alpha}{(c_\alpha + 1)^\beta}\left(\frac{B_1 - (1 - w_2)(\max p_1 - B_1)}{r_1} - 1\right)^\beta\right).$$

From weak altruism, weak efficiency concerns, and satisfiability we can conclude that $\frac{d\delta^-}{dt} \geq 0$. Thus, we have $\delta^-(\Gamma) \leq \delta^-(\Gamma')$, implying that weakly more regular dictators with $\alpha t^\beta > 1$ choose the altruistic corner solution in Γ' compared to Γ.

Step 5 Reluctant sharers When an outside option is introduced, some initial givers switch to that option while the behavior of dictators who sort into the game stays unaffected.

Consider two games $\Gamma = \langle B_1, B_2, P_1, t\rangle$ and $\Gamma' = \langle B_1, B_2, P_1', t\rangle$ with $B_1 > 0$, $B_2 = 0$, $P_1 = [0, B_1]$, and $P_1' = \{[0, B_1], \tilde{p}_1\}$ where $0.5B_1 < \tilde{p}_1 \leq B_1$, i.e. game Γ' is generated from game Γ by adding an outside option to the choice set of the dictator.

Since the two games differ only in the choice set of the dictator, which is equivalent in both games except for the extra outside option in game Γ', the utility functions of a regular dictator in Γ and Γ' are equivalent where the two choice sets overlap. Furthermore, since the dictator's information is not manipulated by the choice of the outside option, her reference point stays the same for the choice of the outside option. We have $r_1(\Gamma) = r_1(\Gamma') =: r_1$ with $r_1 = w_1 B_1$. However, since the outside option leaves the recipient completely uninformed about the choice of the dictator and the rules of the game, his reference point is zero for the outside option choice. We thus have for the reference point of the recipient $r_2(\Gamma) = r_2(\Gamma') =: r_2$ with

$$r_2 = \begin{cases} t w_2 B_1 & \text{if } p_1 \in [0, B_1] \\ 0 & \text{if } p_1 = \tilde{p}_1 \end{cases}$$

The utility of a regular dictator if she chooses the outside option is then given by

$$u(\tilde{p}_1) = \begin{cases} \frac{1}{\beta}(\tilde{p}_1 - w_1 B_1)^\beta & \text{if } \tilde{p}_1 \geq w_1 B_1 \\ -\frac{\delta}{\beta}(w_1 B_1 - \tilde{p}_1)^\beta & \text{if } \tilde{p}_1 < w_1 B_1. \end{cases}$$

Since as noted above the utility functions of a regular dictator in Γ and Γ' are equivalent for

$p_1 \in [0, B_1]$ we have $p_1^+(\Gamma) = p_1^+(\Gamma') =: p_1^+$ with

$$p_1^+ = \frac{1 + c_\alpha w_1 - w_2}{c_\alpha + 1} B_1.$$

and $\delta^+(\Gamma) = \delta^+(\Gamma') =: \delta^+$ with

$$\delta^+ = c_\alpha^{\beta-1} \left(\left(\frac{(1-w_1)B_1}{w_2} \right)^\beta - (c_\alpha + 1)^{1-\beta} \left(\frac{(1-w_1)B_1}{w_2} - 1 \right)^\beta \right).$$

Note first, that no regular dictator with $w_1 > \tilde{p}_1/B_1$ chooses the outside option. By satisfiability, such a dictator can always choose $p_1 \in [0, B_1]$ such that $p_1 \geq r_1$ and $p_2(p_1) \geq r_2$. This yields utility $u(p_1) = (p_1 - r_1)^\beta/\beta + \alpha(t(B_1 - p_1) - r_2)^\beta/\beta \geq 0$, where the inequality follows from weak efficiency concerns. Choosing $p_1' = \tilde{p}_1$ instead yields $u(\tilde{p}_1) = -\delta(w_1 B_1 - \tilde{p}_1)^\beta < 0$. In the following we restrict attention to dictators with $w_1 \leq \tilde{p}_1/B_1$. We have $u(p_1^+) < u(\tilde{p}_1)$ iff

$$\tilde{p}_1 > B_1 \left((c_\alpha + 1)^{\frac{1-\beta}{\beta}} (1 - w_1 - w_2) + w_1 \right) =: \tilde{p}_1^{min}$$

We show that for any Γ and Γ' there exist regular dictators with $\delta \geq \delta^+$ and $\tilde{p}_1^{min} < \tilde{p}_1$, i.e. regular dictators who choose the interior solution in Γ and the outside option in Γ'. For any transfer rate t specified by Γ and Γ' we can find (α, β) satisfying weak altruism and weak efficiency concerns such that $c_\alpha^{1-\beta} \leq 1$. Given such (α, β), for any dictator endowment B_1 specified by Γ and Γ' and any outside option payment $\tilde{p}_1$ specified by Γ' we have $\tilde{p}_1^{min} = 0.5 B_1$ for $w_1 = w_2 = 0.5$. Thus, by continuity of $\tilde{p}_1^{min}$ we can for any Γ and Γ' find (w_1, w_2) in accordance with satisfiability such that $\tilde{p}_1^{min} < B_1$. Since there is no upper bound on the loss aversion parameter of regular dictators, given such (w_1, w_2) there always exist regular dictators with $\delta \geq \delta^+$.

Step 6 Social pressure givers Ceteris paribus, higher susceptibility to social pressure implies higher recipient payoffs at the interior solution but also a higher propensity to choose the outside option in a sorting game.

Higher susceptibility to social pressure corresponds to a higher weight on the opponent's endowment in the reference points, i.e. a higher w_2. We have

$$\frac{\partial p_1^+}{\partial w_2} = -\frac{t}{c_\alpha + t} B_1 < 0 \quad \text{and} \quad \frac{\partial \tilde{p}_1^{min}}{\partial w_2} = -(c_\alpha + 1)^{\frac{1-\beta}{\beta}} B_1 \leq 0$$

where the inequalities follow from weak altruism and weak efficiency concerns.

□

C Details of the econometric specification

Technical details We estimate all parameters by maximum likelihood, and in each case, the likelihood is maximized by a combination of two algorithms: first, using the robust (gradient-free) NEWUOA algorithm (Powell, 2006; Auger et al., 2009), secondly a Newton-Raphson method to ensure convergence. In addition, we cross-test globality of the maxima using a large number of informed starting values. These starting values are derived from estimates for related models on the same data set or from the same model on other data sets. Since we estimated the same model on many different data sets and related models on the same data sets, we were able to generate many informed starting vectors helpful in examining globality of maxima via cross-testing. As is well-known from numerical non-linear maximization (see e.g. McCullough and Vinod, 2003), generating informed starting values is necessary to ensure global optimality, and it proved extremely helpful also in our case. We stopped cross-testing and generating new starting values once

the estimates had converged across all optimization problems simultaneously, based on which we conclude that we approximated the global maxima.

We evaluate significance of differences between models using the Schennach-Wilhelm likelihood ratio test (Schennach and Wilhelm, 2016). This test is robust to both misspecification and arbitrary nesting of models, which is required to allow for the possibility that all models are misspecified and to acknowledge that the nesting structure at least out-of-sample is not necessarily well-defined. In addition, the Schennach-Wilhelm test allows us cluster at the subject level and to thus account for the panel character of the data. We indicate significance of differences between models distinguishing the conventional level of 0.05 and the higher level of 0.01, which roughly implements the Bonferroni correction given four types of dictator game experiments we examine.

As many other experiments involving choice of numbers, responses in dictator games exhibit pronounced round-number patterns. We control for those using the focal choice adjusted logit model, exactly as derived and applied in Breitmoser (2017). The basic idea is that the roundedness of the number to be entered (to choose a given option) determine its "relative focality", which is captured by a focality index $\phi : X \to \mathbb{R}$. The idea that focality is a choice-relevant attribute of options next to utility follows from Gul and Pesendorfer (2001), and given standard axioms including positivity, independence of irrelevant alternatives and narrow bracketing, this implies a generalized logit model of the form

$$\Pr(x) = \frac{\exp\{\lambda u(x) + \kappa\phi(x)\}}{\sum_{x'} \exp\{\lambda u(x') + \kappa\phi(x')\}}. \tag{26}$$

This approach effectively captures round-number effects in stochastic choice, and in turn, simply ignoring the round-number effects as pronounced as in Dictator games was shown to yield substantially biased results in Breitmoser (2017).[17] To avoid spending any degree of freedom here, we use the same focality index as Breitmoser (2017)[18] and set κ equal to 0.8. Robustness checks on both choices are reported in Appendix C.

Capturing heterogeneity One of the more robust finding in behavioral economics is that subjects differ: They have heterogeneous preferences and differing precision in maximizing their preferences, and in addition, we suspect, they also have idiosyncratic reference points. Across subjects, these behavioral primitives are likely correlated. For example, a negative exponent β in the CES utility function implies a flat utility function, and thus to maintain "average precision" in maximizing utility a larger logit-parameter λ is required. Hence, β and λ generally are negatively correlated. For a related observation in the context of risk aversion, see for example Wilcox (2008). The correlation structure itself is unknown, however, and in addition, functional form assumptions about the marginal distributions of parameters seem to be equally difficult to make in the present context. We have only little knowledge about the distribution of individual preferences in generalized dictator games, except that the altruism weight α is likely truncated at say $(-0.5, 0.5)$, and that the exponent β does not seem to comply with a simple continuous distribution (for example, Andreoni and Miller, 2002, estimate that some subjects have linear preferences with β close to 1, some have Cobb-Douglas with $\beta \approx 0$, and others are Leontief with $\beta \to -\infty$).

While somewhat adequate approximations exist for each of these issues, we chose to tackle heterogeneity in a non-parametric manner attempting to combine the strengths of continuous distributions ("random coefficients") and the generality of finite-mixture models (see e.g. McLachlan and Peel, 2004). In a first step, we estimate for each subject the model parameters (preferences α, β,

[17] For example, in the experiment of Korenok et al. (2014), subjects mostly picked multiples of five, typically from option sets ranging from 0 to 20. The most pronounced interior mass points are at choosing payoffs of 10 for both, dictator and recipient. Estimating the reference points of subjects in this experiments without controlling for round number effects yields estimates of reference point 10 each, and in this case, the reference point simply helps to capture the round-number effect. Controlling for the round-number effects, the overall model fit improves drastically and less round-number inspired reference points (deviating from 10 each) are estimated.

[18] That is, multiples of 100 have focality level $\phi_x = 4$, other multiples of 50 have level 3, other multiples of 10 have level 2, other multiples of 5 have level 1, other integers have level 0, other multiples of 0.5 have level -1 and so on. The results are invariant to positive affine transformations of ϕ, i.e. shifting the level of or scaling ϕ does not affect the results.

precision λ, and reference point weights w_1, w_2) individually by maximum likelihood.[19] Then, for the predictions that most of our results rely on, we implement a finite mixture approach where each of the n subjects available in-sample has weight $1/n$ out-of-sample. That is, we model the out-of-sample subject pool to be characterized as a finite mixture of n components, each with prior weight $1/n$, where each component corresponds with one subject from the in-sample data set. For illustration, there are 106 subjects in KMR14. The in-sample estimation yields 106 parameter vectors denoted as $(p_1, p_2, \dots)$. This means that the prediction for the other experiments is that with probability $1/106$ a subject has vector p_1, with probability $1/106$ vector p_2 applies, and so on.

The main advantage of this approach that it allows us to capture distributions of parameters and their correlations without parametric assumptions. Any single parameter estimate is somewhat noisy, obviously, but since maximum likelihood estimates are approximately normally distributed, the errors overall cancel out and we obtain a fairly general description of the joint distribution of the individual parameters. The observed reliability of our out-of-sample predictions corroborates this approach. Finally, the approach is equally applicable to all models, also to the models accounting for say warm glow and cold prickle, or envy and guilt, and in this way it allows for an equally general treatment of heterogeneity across models.

Finally, to adjust for the differences in budgets between experiments and the (potential) differences in the weights of round numbers resulting from the differences in options sets, we allow all individual precision parameters λ and the round-number weight κ to be adjusted jointly across subjects when making predictions between experiments. These two scaling parameters are estimated from the data, but this rescaling is applied equally for all models and does therefore not affect the relative ranking. The likelihood-ratio tests of predictive adequacy also follow Schennach and Wilhelm (2016) as described above.

[19] For numerical reasons, this step is split up into two substeps. First, we estimate individual preference and precision parameters for all reference point weights satisfying $w_1 \geq w_2$ on a grid of step-size 0.1. Secondly, we determine for each individual the likelihood maximizing reference point weights, taking the "smallest" reference point weights in cases of non-uniqueness (non-uniqueness occurs mainly for subjects consistently maximizing their pecuniary payoffs).

Table 4: Instructions differ in the declaration and strength of assignment of endowments

Experiment	Instructions	Classification
AM02	"[...] you are asked to make a series of choices about how to divide a set of tokens between yourself and one other subject in the room."	neutral
HJ06	"[...] you are asked to make a series of choices about how to divide points between yourself and one other subject in the other room"	neutral
CHST07	"[...] you must decide how you want to divide the joint production between yourself and your opponent. In the example above the contributions of the two players to the joint production are 800 NOK and 200 NOK, respectively."	loaded
KMR12	"The blue player has to decide how much of $Y, a fixed amount of money, to pass to the green player and how much to keep for himself/herself. [...] In addition to the money passed by the blue player, the green player will also earn $X."	loaded
KMR13	"Blue will be asked to make a series of 18 choices about how to divide a set of tokens between herself and the Green player. [...] Each choice that Blue makes is similar to the following: Green has 15 points. Divide 50 tokens: HOLD [blank] @ 1 point(s) each, and PASS [blank] @ 2 point(s) each."	neutral (dictator) loaded (recipient)
List07	"Everyone in Room A and in Room B has been allocated $5. The person in Room A (YOU) has been provisionally allocated an additional $5. Participants in Room B have not been allocated this additional $5.[...] decide what portion, if any, of this $5 to transfer to the person you are paired with in Room B. You can also transfer a negative amount: i.e., you can take up to $1 from the person in Room B."	loaded
Bard08	"Each of you has been given £6. [...] You can either leave payments unchanged, increase your own, by decreasing the other person's payment, or decrease your own, increasing the other person's payment."	loaded
KMR14	"In different scenarios you will decide what portion of your endowment to transfer to another participant in the room. Each scenario specifies how much money is in your endowment, how much money is in the OTHER endowment and the range of allowable transfers. In some scenarios you can also transfer a negative amount: i.e., you can take some of the OTHER endowment."	loaded
LMW12	"You will have to decide how to distribute €10 between yourself and the person."	neutral

D Robustness checks in the econometric analysis

The purpose of this section is to show that the results are highly robust to variations in the three econometric assumptions: functional form for reference points (Assumption 4), relative focality of the numbers that may be entered (Footnote 18), extent of round-number effects ($\kappa = 0.8$ in Eq. (26)).

Result 4 (Summary of the robustness checks)**.**

- *We examine four different specifications clarifying how reference points change across contexts (see Definitions 6–8). In line with the theoretical prediction that fairness-based altruism improves model adequacy for all reference point specifications, both descriptive and predictive adequacy (in-sample and out-of-sample) improve highly significantly for all specifications. See Table 5, panel "Aggregate".*
- *We examine two alternative specifications for factoring out round-number effects, the results are very similar for all specifications as shown. See Tables 6 and 7 in comparison to Table 5.*
- *Throughout, we allow for* non-linear inequity aversion *as third benchmark model to extend payoff-based CES altruism. This extension fits substantially worse than the standard linear one examined above and hence was not reported in the paper. See the lines "+ Inequity Aversion (nonl)" in all the tables referenced above.*

D.1 Definitions

For clarity, we first repeat the (deliberately simplistic) base model from the main text.

Definition 6 (Fairness-based altruism (base model))**.** In game $\Gamma = \langle B_1, B_2, P_1, t\rangle$, using $w_1, w_2 \in [0,1]$,

$$r_1(\Gamma) = w_1 \cdot B_1 + w_2 \cdot tB_2$$
$$r_2(\Gamma) = w_2 \cdot B_1 + w_1 \cdot tB_2.$$

Our second robustness check is a model similar to Definition 6, but other endowments are weighed by transfer rate. This implicitly yields inequity averse reference points for $w_1 = w_2$ (scaled down or up if $w_1 + w_2 \gtrless 1$). It is equivalent to Definition 6 if $t = 1$. By comparing it to Definition 6, we can evaluate if subjects take the transfer rate into account when forming reference points. Notable special cases are CES ($w_1 = w_2 = 0$), and inequity aversion/egalitarian ($w_1 = w_2 = 0.5$), strict libertarian ref points ($w_1 = 1, w_2 = 0$). Obviously, the model allows for a continuum in-between.

Definition 7 (Fairness-based altruism 2 (robustness check I))**.** In game $\Gamma = \langle B_1, B_2, P_1, t\rangle$, using $w_1, w_2 \in [0,1]$,

$$r_1(\Gamma) = w_1 \cdot B_1 + w_2 \cdot B_2$$
$$r_2(\Gamma) = w_2 \cdot tB_1 + w_1 \cdot tB_2.$$

Our second robustness check adapts the base model in Definition 6 by allowing for the background income to equate with the minimal payoff, rather than the outside-laboratory payoff.

Definition 8 (Fairness-based altruism 3 (robustness check II))**.** In game $\Gamma = \langle B_1, B_2, P_1, t\rangle$, using $w_1, w_2 \in [0,1]$,

$$r_1(\Gamma) = \min p_1 + w_1 \cdot (B_1 - \min p_1) + w_2 \cdot (tB_2 - \min p_2)$$
$$r_2(\Gamma) = \min p_2 + w_2 \cdot (B_1 - \min p_1) + w_1 \cdot (tB_2 - \min p_2).$$

Our final robustness check is the arguably most realistic model used in the theoretical analysis, weighing by transfer rate and using the minimal payoff as background income. This model usually fits best. It contains status-quo-based reference points ($w_1 = w_2 = 0$) and strict expectations-based reference points ($w_1 + w_2 = 1$) as the most notable special cases, and by allowing for $w_1 + w_2 \in (0,1)$ all convex combinations are also included.

Definition 9 (Fairness-based altruism 4 (robustness check III))**.** In game $\Gamma = \langle B_1, B_2, P_1, t \rangle$, using $w_1, w_2 \in [0,1]$,

$$r_1(\Gamma) = \min p_1 + w_1 \cdot (B_1 - \min p_1) + w_2 \cdot (B_2 - \min p_2/t),$$
$$r_2(\Gamma) = \min p_2 + w_2 \cdot t \cdot (B_1 - \min p_1) + w_1 \cdot (t \cdot B_2 - \min p_2).$$

As non-linear model of inequity aversion, we use the following straightforward extension of CES altruism.

Definition 10 (Non-linear inequity aversion)**.** Using the notation in the main text, non-linear inequity aversion is defined as follows:

$$u(\pi) = (1 - \alpha_1 - \alpha_2 - \alpha_3) \cdot \pi_1^\beta + \alpha_1 \, \pi_2^\beta - \alpha_2 \cdot |\pi_1 - \pi_2|_+^\beta - \alpha_3 \cdot |\pi_2 - \pi_1|_+^\beta.$$

(+ Inequity Aversion (nonl))

Finally, as simplified focality weights as robustness check for the standard focality weights described above (Footnote 18, which follows Breitmoser (2017)), we use the following.

Definition 11 (Simplified focality weights)**.** All numbers that are multiples of 5 have focality weight $\phi = 1$ in Eq. (26), all other numbers have focality weight $\phi = 0$.

D.2 Results

Table 5: Predictions for standard focality weights and $\kappa = 0.8$ (results from main text)

Calibrated on	Altruism is ...	Descriptive Adequacy	Predictive Adequacy	Details on predictions of ...			
				Dictator	Endowments	Taking	Sorting
Aggregate	Payoff based (CES)	5839.8	27404.1	9343.1	9631.8	5546.8	2882.4
	+ Warm Glow/Cold Prickle	5354.6^{++}	28075^{--}	9896.5^{--}	9581.6	5617.8^{-}	2979.1^{--}
	+ Inequity Aversion	5453.9^{++}	27447.9	9094^{+}	9859.8^{--}	5600.4^{--}	2893.7
	+ Inequity Aversion (nonl)	5718.2^{+}	27435	9196.1	9811.9^{--}	5546.4	2880.6
	Fairness based	5035.7^{++}	26674.4^{++}	9093.2^{++}	9385^{++}	5451^{++}	2745.2^{++}
	Fairness based (adj)	5035.7^{++}	25740.4^{++}	8883.6^{++}	9023.5^{++}	5212.2^{++}	2631.2^{++}
	Fairness based 2	5181.4^{++}	26919.5^{++}	9108.6^{++}	9529.5	5473.3^{+}	2808.2^{++}
	Fairness based 2 (adj)	5181.4^{++}	26209^{++}	8852.9^{++}	9179.9^{++}	5393.2^{++}	2793^{++}
	Fairness based 3	5048.4^{++}	27064.9^{+}	9221.4	9640.7	5494.5^{+}	2708.2^{++}
	Fairness based 3 (adj)	5048.4^{++}	25920^{++}	8559.7^{++}	9306.4^{++}	5393.2^{++}	2670.7^{++}
	Fairness based 4	4936.9^{++}	26945^{++}	9308.3	9354.1^{++}	5493.6^{+}	2789^{++}
	Fairness based 4 (adj)	4936.9^{++}	25703.9^{++}	8594.5^{++}	9167.1^{++}	5286.7^{++}	2665.6^{++}
Dictator games	Payoff based (CES)	1460.9	8950.5	1343.4	4339	2353.3	914.7
	+ Warm Glow/Cold Prickle	1507.3^{--}	8854.6	1343	4218.4^{+}	2375.2	917.9
	+ Inequity Aversion	1234.6^{++}	8794.8^{++}	1217.1^{+}	4311.7	2360.7	905.3
	+ Inequity Aversion (nonl)	1314.9^{++}	8943.8	1271.4^{++}	4391.2^{--}	2357.8	923.3
	Fairness based	1146.6^{++}	8758^{++}	1279.8^{+}	4273.8^{+}	2316.6^{+}	887.7
	Fairness based (adj)	1146.6^{++}	8603.9^{++}	1263.9^{+}	4152.5^{++}	2300.8^{++}	888.2
	Fairness based 2	1146.4^{++}	8849.2^{+}	1276.4^{+}	4355.8	2325^{+}	892^{+}
	Fairness based 2 (adj)	1146.4^{++}	8585.3^{++}	1265.7^{+}	4119.5^{++}	2309.7^{++}	892^{+}
	Fairness based 3	1055^{++}	8818.4^{+}	1272.6^{+}	4336.7	2321.5^{+}	887.5^{+}
	Fairness based 3 (adj)	1055^{++}	8673.6^{++}	1255.2^{+}	4231.6	2307.8^{+}	880.5^{+}
	Fairness based 4	1050.9^{++}	8715.2^{++}	1268.8^{+}	4240.1^{++}	2324.2	882.1^{++}
	Fairness based 4 (adj)	1050.9^{++}	8662.1^{++}	1252.5^{+}	4219.8^{++}	2309.4^{+}	881.9^{++}
Gen Endowments	Payoff based (CES)	2896.6	8752.9	4260.4	826.1	2613.8	1052.7
	+ Warm Glow/Cold Prickle	2395.5^{++}	8967.8^{--}	4289.6	954.5^{--}	2649.7	1074
	+ Inequity Aversion	2800.1^{+}	8916.4^{--}	4333.6^{-}	849.9	2663^{--}	1069.9^{--}
	+ Inequity Aversion (nonl)	2923.3	8703.6^{+}	4235.2	824.5	2599.8^{+}	1044^{++}
	Fairness based	2662.7^{++}	8416.7^{++}	4084.2^{++}	767.9^{+}	2565.9^{+}	998.7^{++}
	Fairness based (adj)	2662.7^{++}	7867.7^{++}	3985.8^{++}	637.1^{++}	2351^{++}	895.4^{++}
	Fairness based 2	2769.6^{++}	8615.1^{++}	4157.7^{+}	819.4	2580.2	1057.8
	Fairness based 2 (adj)	2769.6^{++}	8312.5^{++}	3995.9^{++}	751.5^{++}	2521.6^{++}	1045.1
	Fairness based 3	2730^{++}	8626.1^{++}	4236.6	822.1	2606.2	961.2^{++}
	Fairness based 3 (adj)	2730^{++}	7928.2^{++}	3692.7^{++}	778.5^{+}	2524.5^{++}	934^{++}
	Fairness based 4	2662.7^{++}	8754.3	4319.1^{-}	782	2601.1	1052
	Fairness based 4 (adj)	2662.7^{++}	7710.2^{++}	3719.7^{++}	643.3^{++}	2413.4^{++}	935.2^{++}
Taking Games	Payoff-based (CES)	1482.4	9700.7	3739.3	4466.7	579.7	914.9
	+ Warm Glow/Cold Prickle	1451.8	10252.5^{--}	4263.8^{--}	4408.7	592.8	987.2^{--}
	+ Inequity Aversion	1419.2^{+}	9736.7	3543.3^{++}	4698.2^{--}	576.6	918.5
	+ Inequity Aversion (nonl)	1479.9	9787.7	3689.5	4596.1^{--}	588.8^{-}	913.3
	Fairness based	1226.4^{++}	9499.7^{+}	3729.2	4343.2	568.5^{+}	858.8^{++}
	Fairness based (adj)	1226.4^{++}	9270.3^{++}	3633^{+}	4232.9^{++}	559.3^{++}	846.6^{++}
	Fairness based 2	1265.5^{++}	9455.3^{++}	3674.5	4354.2^{+}	568.2	858.3^{++}
	Fairness based 2 (adj)	1265.5^{++}	9310.1^{++}	3590.4^{+}	4305.4^{++}	560.9^{+}	854.9^{++}
	Fairness based 3	1263.4^{++}	9620.4	3712.2	4482	566.9	859.4^{++}
	Fairness based 3 (adj)	1263.4^{++}	9312.1^{++}	3603.2^{+}	4295.4^{+}	559.8^{+}	855.2^{++}
	Fairness based 4	1223.4^{++}	9475.5^{++}	3720.4	4332^{+}	568.2^{+}	855^{++}
	Fairness based 4 (adj)	1223.4^{++}	9331.8^{++}	3620.6	4302.4^{+}	562.9^{++}	847.5^{++}

Table 6: Predictions for simplified focality weights and $\kappa = 0.8$ (robustness check)

Calibrated on	Altruism is ...	Descriptive Adequacy	Predictive Adequacy	Details on predictions of ... Dictator	Endowments	Taking	Sorting
Aggregate	Payoff based (CES)	5968.4	27868.5	10084.1	9676.9	5277.3	2830.1
	+ Warm Glow/Cold Prickle	5546.9^{++}	28922.2^{--}	10687^{--}	9846.6^{-}	5428^{--}	2960.7^{--}
	+ Inequity Aversion	5593.9^{++}	27994.9	9944	9772.7	5377.8^{--}	2900.4^{--}
	+ Inequity Aversion (nonl)	6128.5^{--}	28905.7^{--}	10752.5^{--}	9827^{--}	5353.9^{--}	2972.4^{--}
	Fairness based	4677.3^{++}	27288.5^{++}	9790.8^{++}	9560.8	5232.4^{+}	2704.5^{++}
	Fairness based (adj)	4677.3^{++}	26308.2^{++}	9618.3^{++}	9107.8^{++}	5000.7^{++}	2591.4^{++}
	Fairness based 2	5023.7^{++}	26894.6^{++}	9920.8^{+}	8986.9^{++}	5275.3	2711.6^{++}
	Fairness based 2 (adj)	5023.7^{++}	26240.1^{++}	9659.8^{++}	8759^{++}	5148^{+}	2683.3^{++}
	Fairness based 3	5258^{++}	27031.7^{++}	9843.9^{++}	9180.6^{++}	5270.6	2736.6^{++}
	Fairness based 3 (adj)	5258^{++}	26174.3^{++}	9591.9^{++}	8875.1^{++}	5133.9^{+}	2583.4^{++}
	Fairness based 4	5258^{++}	26772.1^{++}	9733.1^{++}	9174.7^{++}	5202.5^{+}	2661.8^{++}
	Fairness based 4 (adj)	5258^{++}	25472.9^{++}	8880^{++}	8933.2^{++}	5088.7^{++}	2581^{++}
Dictator games	Payoff based (CES)	1697.2	8998.3	1462.4	4387.3	2253.5	895.1
	+ Warm Glow/Cold Prickle	1715.9	9104.6^{--}	1502.8^{--}	4377.3	2304^{--}	920.4^{-}
	+ Inequity Aversion	1390.2^{++}	8834.1^{+}	1352^{+}	4313.9	2268.5	899.7
	+ Inequity Aversion (nonl)	1753^{--}	9117.8^{--}	1484.9^{-}	4418.9	2295.4^{--}	918.6^{-}
	Fairness based	1392^{++}	8807.9^{++}	1396.7^{++}	4335.2	2208.4^{++}	867.5
	Fairness based (adj)	1392^{++}	8473.5^{++}	1349.4^{++}	4080^{++}	2184.7^{++}	861^{+}
	Fairness based 2	1400.9^{++}	8757.5^{++}	1442.9	4170.8^{++}	2258	885.9
	Fairness based 2 (adj)	1400.9^{++}	8654.1^{++}	1437	4090.9^{++}	2248.6	879.1
	Fairness based 3	1392.3^{++}	8801^{++}	1391.2^{++}	4266^{++}	2270.1	873.7
	Fairness based 3 (adj)	1392.3^{++}	8548.6^{++}	1348.2^{++}	4071.4^{++}	2263.4	867.1^{+}
	Fairness based 4	1392.7^{++}	8707.2^{++}	1360.6^{+}	4234.8^{+}	2257.3	854.4
	Fairness based 4 (adj)	1392.7^{++}	8529.2^{++}	1356.5^{+}	4093.5^{++}	2241.9	838.8^{+}
Gen Endowments	Payoff based (CES)	2870.3	8828.2	4503	840.8	2441.2	1043.2
	+ Warm Glow/Cold Prickle	2438.7^{++}	9018.4^{--}	4518.8	936.9^{--}	2500.4^{--}	1062.4
	+ Inequity Aversion	2837.6	9057.8^{--}	4650.6^{--}	841.7	2504.8^{--}	1060.6^{-}
	+ Inequity Aversion (nonl)	2926.5^{--}	9003.2^{--}	4609.8^{--}	871.4^{--}	2453.2	1068.8^{--}
	Fairness based	2149.8^{++}	8650.6^{++}	4372.5^{++}	836.2	2448.7	993.1^{+}
	Fairness based (adj)	2149.8^{++}	8159.9^{++}	4308.6^{++}	703.3^{++}	2250.8^{++}	898.8^{++}
	Fairness based 2	2387^{++}	8763.2	4561.1	767^{++}	2443.9	991.2^{+}
	Fairness based 2 (adj)	2387^{++}	8321.2^{++}	4347.6^{+}	676.6^{++}	2329.4^{+}	969.1^{++}
	Fairness based 3	2636.8^{++}	8763.8	4544	762.5^{++}	2427.8	1029.4
	Fairness based 3 (adj)	2636.8^{++}	8216^{++}	4362.4^{++}	673^{++}	2299.8^{++}	882.2^{++}
	Fairness based 4	2586.3^{++}	8494.9^{++}	4401^{++}	774.9^{++}	2366^{++}	953^{+}
	Fairness based 4 (adj)	2586.3^{++}	7618.3^{++}	3757.5^{++}	696.4^{++}	2275.2^{++}	890.7^{++}
Taking Games	Payoff-based (CES)	1400.9	10041.9	4118.7	4448.7	582.6	891.9
	+ Warm Glow/Cold Prickle	1392.3	10799.2^{--}	4665.4^{--}	4532.4^{-}	623.5^{--}	977.9^{--}
	+ Inequity Aversion	1366.1^{+}	10103	3941.3^{+}	4617^{--}	604.6^{--}	940.1^{--}
	+ Inequity Aversion (nonl)	1448.9^{--}	10784.7^{--}	4657.8^{--}	4536.7^{--}	605.3^{--}	985^{--}
	Fairness based	1135.5^{++}	9830^{+}	4021.5	4389.4	575.2	843.9^{+}
	Fairness based (adj)	1135.5^{++}	9676.3^{++}	3959.4^{++}	4323.5^{+}	564.2^{++}	830.7^{++}
	Fairness based 2	1235.8^{++}	9374^{++}	3916.9^{++}	4049.1^{++}	573.4	834.6^{++}
	Fairness based 2 (adj)	1235.8^{++}	9266.3^{++}	3874.2^{++}	3990.5^{++}	569^{+}	834.1^{++}
	Fairness based 3	1228.9^{++}	9466.8^{++}	3908.7^{++}	4152.1^{++}	572.7	833.4^{++}
	Fairness based 3 (adj)	1228.9^{++}	9411.2^{++}	3880.3^{++}	4129.7^{++}	569.6^{+}	833^{++}
	Fairness based 4	1279.1^{++}	9569.9^{++}	3971.4	4164.9^{++}	579.1	854.4
	Fairness based 4 (adj)	1279.1^{++}	9326.9^{++}	3765.1^{++}	4142.3^{++}	570.5	850.5^{+}

Table 7: Predictions for standard focality weights and $\kappa = 0.6$ (robustness check)

Calibrated on	Altruism is ...	Descriptive Adequacy	Predictive Adequacy	Details on predictions of ...			
				Dictator	Endowments	Taking	Sorting
Aggregate	Payoff based (CES)	5858.5	27706.5	9385.7	9895.7	5561.1	2864.1
	+ Warm Glow/Cold Prickle	5385.4^{++}	28483.9^{--}	10056.6^{--}	9809.1	5639^{-}	2979.2^{--}
	+ Inequity Aversion	5458.5^{++}	27412.6^{+}	9048^{+}	9844.8	5636.8^{--}	2882.9
	+ Inequity Aversion (nonl)	5703.4^{++}	27412.1^{+}	9166.5^{+}	9824.5	5548.6	2872.5
	Fairness based	5030.7^{++}	26719^{++}	9156.4^{+}	9359.3^{++}	5480.2^{+}	2723.1^{++}
	Fairness based (adj)	5030.7^{++}	25647.1^{++}	8894.5^{++}	8962.8^{++}	5193.7^{++}	2606.1^{++}
	Fairness based 2	5175.3^{++}	27115^{++}	9350.9	9488.8^{++}	5487.3^{+}	2788^{++}
	Fairness based 2 (adj)	5175.3^{++}	26237.4^{++}	9054^{++}	9037.9^{++}	5402.6^{++}	2752.8^{++}
	Fairness based 3	5015.1^{++}	26985.5^{++}	9207.2	9573.8^{+}	5505.2	2699.3^{++}
	Fairness based 3 (adj)	5015.1^{++}	25725.8^{++}	8509.8^{++}	9191.6^{++}	5401.6^{++}	2632.9^{++}
	Fairness based 4	4927.3^{++}	26759.6^{++}	9189.9	9334.5^{++}	5527	2708.2^{++}
	Fairness based 4 (adj)	4927.3^{++}	25558^{++}	8503.9^{++}	9130.5^{++}	5279.5^{++}	2654.1^{++}
Dictator games	Payoff based (CES)	1493.5	9087.2	1374.2	4442.4	2370.4	900.3
	+ Warm Glow/Cold Prickle	1533	9012.6	1369.2	4355.5^{+}	2378	910
	+ Inequity Aversion	1238.4^{++}	8835.5^{++}	1204.7^{++}	4341.1^{+}	2386.6	903
	+ Inequity Aversion (nonl)	1326.8^{++}	8999.9	1245.1^{++}	4472.8	2366.4	915.6^{-}
	Fairness based	1165.2^{++}	8725.1^{++}	1278^{++}	4256.8^{++}	2317.2^{+}	873.1
	Fairness based (adj)	1165.2^{++}	8486.5^{++}	1235.9^{++}	4077.4^{++}	2302.2^{++}	872.4
	Fairness based 2	1168.9^{++}	8738.7^{++}	1285.8^{++}	4245.4^{++}	2334.2^{+}	873.3
	Fairness based 2 (adj)	1168.9^{++}	8580.8^{++}	1256^{++}	4133.4^{++}	2321.1^{+}	871.8^{+}
	Fairness based 3	1066.6^{++}	8756.4^{++}	1270.3^{+}	4286.8^{+}	2321.5^{+}	877.7^{+}
	Fairness based 3 (adj)	1066.6^{++}	8556.3^{++}	1247.2^{++}	4124.1^{++}	2310^{++}	876.4^{+}
	Fairness based 4	1066.7^{++}	8690.2^{++}	1261.5^{++}	4225.2^{++}	2332.5	870.9^{+}
	Fairness based 4 (adj)	1066.7^{++}	8580.2^{++}	1238.9^{++}	4161.1^{++}	2312.8^{+}	868.9^{+}
Gen Endowments	Payoff based (CES)	2867	8696	4197.9	829.2	2613.8	1055.2
	+ Warm Glow/Cold Prickle	2383.2^{++}	9015.5^{--}	4311.2^{--}	961.3^{--}	2668	1075.1
	+ Inequity Aversion	2791.6^{+}	8899.2^{--}	4291.5^{--}	855.6	2681.2^{--}	1070.9^{-}
	+ Inequity Aversion (nonl)	2892	8677.4	4249.7	786.3^{++}	2595.8	1045.7
	Fairness based	2631.6^{++}	8479.8^{++}	4122.8^{+}	769.8^{+}	2586.6	1000.7^{++}
	Fairness based (adj)	2631.6^{++}	7884.6^{++}	4026^{++}	640.5^{++}	2329.2^{++}	890.3^{++}
	Fairness based 2	2731.8^{++}	8809.8^{--}	4348.1^{--}	813.5	2591	1057.2
	Fairness based 2 (adj)	2731.8^{++}	8468.9^{++}	4154	748.4^{++}	2522.8^{++}	1045
	Fairness based 3	2673.4^{++}	8750.8	4315.3^{--}	848.7	2621.3	965.5^{++}
	Fairness based 3 (adj)	2673.4^{++}	7915.3^{++}	3677^{++}	788.2^{+}	2532.8^{+}	918.8^{++}
	Fairness based 4	2626.2^{++}	8624.4	4242.4	776.7^{+}	2618.6	986.7^{++}
	Fairness based 4 (adj)	2626.2^{++}	7705.4^{++}	3715.2^{++}	647^{++}	2403.3^{++}	941.4^{++}
Taking Games	Payoff-based (CES)	1498.1	9923.3	3813.6	4624.1	576.9	908.6
	+ Warm Glow/Cold Prickle	1469.1	10455.7^{--}	4376.2^{--}	4492.4^{++}	593^{-}	994.1^{--}
	+ Inequity Aversion	1428.5^{+}	9677.9^{++}	3551.8^{++}	4648.1	568.9^{+}	909
	+ Inequity Aversion (nonl)	1484.7	9734.8^{+}	3671.7^{+}	4565.4	586.4^{-}	911.2
	Fairness based	1234^{++}	9514.1^{++}	3755.7	4332.7^{++}	576.5	849.3^{++}
	Fairness based (adj)	1234^{++}	9277.5^{++}	3631.6^{++}	4243.8^{++}	561.3^{++}	842.4^{++}
	Fairness based 2	1274.6^{++}	9566.5^{++}	3717	4429.9^{++}	562.2	857.5^{++}
	Fairness based 2 (adj)	1274.6^{++}	9187.6^{++}	3643^{++}	4153.4^{++}	557.7^{+}	835^{++}
	Fairness based 3	1275.1^{++}	9478.3^{++}	3621.6^{++}	4438.3^{+}	562.3	856.1^{++}
	Fairness based 3 (adj)	1275.1^{++}	9255.1^{++}	3584.5^{++}	4278.2^{++}	557.2^{+}	836.7^{++}
	Fairness based 4	1234.4^{++}	9445.1^{++}	3686	4332.5^{++}	575.9	850.7^{++}
	Fairness based 4 (adj)	1234.4^{++}	9273.4^{++}	3548.9^{++}	4321.1^{++}	562.2^{++}	842.8^{++}