

# AXIOMATIC FOUNDATIONS FOR FAIRNESS-MOTIVATED PREFERENCES

MARTIN EILIV SANDBU

Forthcoming in *Social Choice and Welfare*

**ABSTRACT.** Much work in social choice theory takes individual preferences as uninvestigated inputs into aggregation functions designed to reflect considerations of fairness. Advances in experimental and behavioural economics show that fairness can also be an important motivation in the preferences of individuals themselves. A proper characterisation of how fairness concerns enter such preferences can enrich the informational basis of many social choice exercises.

This paper proposes axiomatic foundations for individual fairness-motivated preferences that cover most of the models developed to rationalise observed behaviour in experiments. These models fall into two classes: Outcome-based models, which see preferences as defined only over distributive outcomes, and context-dependent models, which allow rankings over distributive outcomes to change systematically with non-outcome factors. I accommodate outcome-based and context-sensitive fairness concerns by modelling fairness-motivated preferences as a *reference-dependent preference structure*. I first present a set of axioms and two theorems that generate commonly used outcome-based models as special cases. I then generalise the axiomatic basis to allow for reference-dependence, and derive a simple functional form in which the weight on each person's payoff depends on a reference vector of how much each person deserves.

*Acknowledgement.* This paper is based on Chapter 2 of my Ph. D. dissertation (Sandbu 2003), which I wrote with support from the Research Council of Norway. I am particularly grateful to Jerry Green for encouraging me to pursue the axiomatic approach presented here. For helpful comments, critiques, and suggestions, I thank Christopher Avery, Walter Bossert, Federico Girosi, David Laibson, Xaq Pitkow, Michael Sandel, Amartya Sen, Richard Tuck, Richard Zeckhauser, seminar participants at Harvard University and Université de Montréal, and several anonymous referees, including one who identified an error in an earlier version.

## 1. INTRODUCTION

Theories of social choice typically proceed by taking individual preferences over states of the world as inputs, and applying normatively desirable conditions on how they are to affect choices for the society as a whole. Usually the individual preferences themselves are left uninvestigated, except perhaps for certain formal properties such as single-peakedness that may make aggregation easier. The task is to find reasonable rules for compromising between individuals whose preferences differ, while withholding judgment as much as possible on the preferences themselves. A division of labour assigns analysis of individual choice and preference to decision theory, and trade-offs between those preferences to social choice theory and welfare economics.

Recent research in behavioural economics, however, has shown that *individual* preferences incorporate rich and complex judgments about how to allocate payoffs across different people. The stylised facts include that people often make positive gifts in Dictator games (Forsythe *et al.* 1994, and others), refuse very unequal allocations in Ultimatum games (Güth *et al.* 1982, Roth *et al.* 1991, and others), and contribute to public goods (Ledyard 1995, Fehr & Schmidt 2001). Another set of observations shows that behaviour is very sensitive to contextual features of the games that should not matter to forward-looking or “consequentialist” agents.<sup>1</sup> Such findings suggest that at least some components of individuals’ preferences are motivated by their own theories of fairness and of how social choices should be made.

These components should be of interest to social choice theories, for at least two reasons. First, when we use individual preferences as inputs into a normatively acceptable social choice, we may at a minimum want to differentiate between the component of individual preferences motivated by a view of fairness, and the components not thus motivated. If we do not, we risk “punishing” those whose preferences deviate from their own interests for reasons of morality and justice, by treating such preferences as equivalent to the self-interested preferences of more unscrupulous people. Second, we may want to respect the specific theories of fairness that shape people’s distributive

---

<sup>1</sup>For example, the likelihood that a Responder in an Ultimatum game will reject an unequal offer varies systematically with the alternative offer available to the Proposer (Güth *et al.* 2001). More generally, strategically identical choice situations elicit different behaviour depending on the supergame they are embedded in (Prasnikar & Roth 1992, Andreoni *et al.* 2002, Camerer 2003, chapter 2) and on the characteristics of the other players, even when this information is strategically irrelevant for the outcome in terms of final payoffs.

preferences. If we want social choice rules and social welfare functions to track the preferences of individuals, it seems arbitrary not to want the axioms and principles that constitute those rules and functions to track the equivalent axioms and principles that characterise individual preferences, if such can be found.<sup>2</sup>

The first step towards enriching the informational basis of social choice theory in this way must be an exercise in empirically informed decision theory, designed to characterise the ways in which individuals might conceptualise fairness. This paper makes a contribution towards that aim. I propose a set of axiomatic characteristics of the kind of individual preferences that have been documented in the experimental literature. Thus my axioms are mostly rationalised not on ethical grounds, as is normal in social choice theory, but rather on positive grounds. I will nevertheless discuss briefly to what extent the axioms are appealing on normative grounds—from perspectives of rationality or of ethics. The idea is that if we know which basic principles *in fact* govern people’s distributive choices, we can at a later stage assess what normative role those principles ought to play in a social choice perspective. This article, however, focuses on the first stage, which is to characterise the ways in which individual preferences incorporate fairness concerns.

## 2. CONCEPTUAL AND FORMAL PRELIMINARIES

**2.1. Two ways in which fairness concerns enter preferences.** Theories of fairness-motivated preferences can be categorised into two classes: Outcome-based and non-outcome-based models. In the former, only ultimate outcomes (what Amartya Sen (1997, 1999) calls “culmination outcomes”) matter, but unlike in conventional theory, the relevant description of the outcomes—the objects of preference—are payoff vectors rather than only the decision-maker’s payoff. Each outcome-based model postulates that agents care about certain distributive characteristics of the payoff vectors—such as equality or efficiency—and specify a trade-off between distributive concerns and between those concerns and the desire for selfish gain. Important outcome-based investigations include Fehr and Schmidt’s (1999) and Bolton and Ockenfels’s (2000) models of inequality

---

<sup>2</sup>I focus on individuals’ concern for fairness, but there are other aspects of individual preferences that are of relevance to social choice theory. One example is how preferences may be affected by deliberation—the claim has been made that post-deliberative, “laundered” preferences should be favoured in social choice aggregation (Dryzek & List 2003).

aversion; Charness and Rabin's (2002) models of utilitarian "social welfare preferences" or combined utilitarian and "maximin" preferences; relative income or "catching up with the Joneses"-preferences (Ok & Koçkesen 2000); and Andreoni and Millers's (2002) demonstration that the altruistic behaviour in dictator games can be rationalised by a preference relation over payoff vectors obeying the standard axioms of revealed preference.

In non-outcome-based theories, factors other than the ultimate payoff allocations are allowed to influence preferences. Sen (1995, 1997, 1999), for example, draws attention to the *act of choice*. He argues that the value of an outcome may depend on *who* makes the choice that brings it about (chooser-dependence) and what the *alternative* choices are (menu-dependence). A specific theory of how the acts of choice matter in the context of fairness is *reciprocity theory* (Rabin 1993, Falk & Fischbacher 2000, Dufwenberg & Kirchsteiger 2004), which postulates that people want to reward kind (or fair) intentions and punish mean (or unfair) intentions. Since intentions are revealed through actions, reciprocity motives make preferences sensitive to the history of interaction. More generally, it has long been recognised that fairness may be an attribute not just of distributive outcomes, but of the process that brings them about (Rawls 1971, Nozick 1974). Empirically, judgments of "procedural justice" have been shown to be powerful determinants of behaviour and reported well-being in a wide range of economic situations (Frey *et al.* 2004 provide an overview).<sup>3</sup>

This paper treats the non-outcome factors as affecting a decision-maker subjective perception of how much different individuals *deserve*. The idea is that outcome-based distributive preferences apply in the special case where what everyone deserves (in the decision-maker's mind) is fixed, but that in general a variety of factors can change people's deservingness and therefore change the decision-maker's view of how their respective payoffs ought to be traded off against each other. For example, if reciprocity theory is correct, then one relevant non-outcome factor is how much somebody has helped me in the past; those who have helped me deserve more than those who have not. Similarly, somebody who has lied to me may deserve less than somebody who has not. I will not propose a theory of how any given non-outcome factor affects deservingness

---

<sup>3</sup>A large number of non-outcome factors have been shown to influence choices over payoff vectors. They include strategically irrelevant parts of the set of available outcomes (Andreoni *et al.* 2002, Sandbu 2007), promising (Charness & Dufwenberg 2006), truth-telling (Brandts & Charness 1999), status (Cox *et al.* 2007), gender (Andreoni & Petrie 2004), and property rights (Gächter & Riedl 2005).

(there would have to be one theory for each factor); my concern is to axiomatise how judgments of deservingness affect ultimate preferences over payoff distributions. To do so, I assume that non-outcome factors collectively determine a “fair reference allocation” (a vector denoting each individual’s deserved payoff), and model fairness-motivated preferences as *reference-dependent preference structures*, i.e. functional relationships from reference points to preference relations.<sup>4</sup> The purely outcome-based models then constitute the special case where the reference allocation is constant or (formally equivalently) where there is no reference-dependence.

“Deservingness” here does not necessarily refer to the most fair way to divide the *available* sum of money. How much the decision-maker thinks the individuals deserve need not correspond to a feasible and exhaustive division of the available monetary payoffs. In non-constant-sum games, how much is available depends on the players’ choices. But even in zero-sum-games where there is a fixed available amount, she may think that it is impossible to give everyone all that they deserve, or, in contrast, that giving everyone what they deserve will leave some payoff “left over.” Consider reciprocity motives. When you act kindly to me, I may now judge that you deserve more than before, without thinking that *I* therefore deserve less than before. Indeed your act of kindness can cause me some discomfort as it increases the sum of “deservingnesses” without increasing the available pot of money—to give you what you deserve, I have to take less than what I deserve. Similarly, if you are unkind to me, I will judge you less deserving as a result, without thinking that it makes me more deserving than before, except relatively to you. (Nozick (1974) gives a philosophical argument about the logical independence of deservingness and entitlements to shares of available money.) What decision to make in such cases is precisely what an axiomatic approach should aim to formalise.<sup>5</sup>

<sup>4</sup>This view of fairness-motivated preferences allows a conceptual distinction between what the agent considers a fair allocation (what people deserve) and what she actually chooses. These should not be conflated—there is no reason to think that agents always choose what they think is most fair, as self-interest remains a strong motive even for fairness-motivated individuals.

<sup>5</sup>An anonymous referee expresses the following worry: “Without knowing the set of feasible payoff vectors, one can certainly think of who is more [and] less deserving [and] even [...] that individual *i* is twice as deserving as individual *j*. But I am not sure that one can think of what payoffs the different individuals deserve without knowing the feasibility constraints in the specific case”. The examples in the main text, however, shows how it need not be incoherent for the vector of deservingness not to correspond to an actually feasible payoff allocation. Whether it reasonable is another matter. The reviewer’s worry may be that it is only reasonable to judge *relative* deservingness, that is, how deserving one individual is compared to others. This is the kind of intuition that should guide our choice of axioms. This

The second half of this section introduces the notation for the formalisation. Section 3 shows how a set of plausible axioms generates outcome-based models of fairness preferences as special cases of reference-*independence*. Section 4 generalises the axiomatic basis to derive a reference-dependent utility function over payoff vectors that can encompass both outcome-based and non-outcome based fairness concerns. Section 5 surveys related literature, and section 6 concludes. All theorems are proved in the appendix.

**2.2. Notation.** Adopting the approach used by Tversky & Kahneman (1991) and Munro & Sugden (2003), we model a decision-maker’s preferences as a *reference-dependent preference structure*, which is a function from the set of reference points to the set of preference relations. We shall consider preferences defined over the set of all  $(n + 1)$ -dimensional vectors with nonnegative components,  $\mathbb{R}_+^{n+1}$ . We interpret its elements as *payoff allocations* among  $n$  individuals plus the decision-maker. Denote the agent whose preferences are under consideration as person 0, let  $N \equiv \{1, \dots, n\}$  be the set of “others,” and  $N_0 \equiv \{0\} \cup N$  the set of all individuals concerned, including the decision-maker. A subset of the concerned individuals will be denoted  $I \subseteq N_0$ , and the components of an allocation  $\mathbf{x}$  corresponding to the individuals in  $I$  is  $\mathbf{x}_I \in \mathbb{R}_+^{|I|}$  (where  $|I|$  is the cardinality of  $I$ ). We denote the complement of  $I$  by  $\sim I$  and the set of all individuals except  $i$ , or  $N_0 \setminus \{i\}$ , by  $\sim i$ . We only consider individuals that are *essential* in the sense of Debreu (1983); that is, an individual is not in  $N_0$  if the decision-maker is *always* indifferent to his or her payoff.

A generic allocation in  $\mathbb{R}_+^{n+1}$  is denoted  $\mathbf{x} = (x_0, x_1, \dots, x_n) \in \mathbb{R}_+^{n+1}$ . A preference *relation* is the binary “at least as good as” relation which ranks the elements of  $\mathbb{R}_+^{n+1}$ . A preference *structure* specifies a preference relation  $\succeq_{\mathbf{r}}$  (to be read “at least as good as, given  $\mathbf{r}$ ”) for each *reference allocation* (or reference point)  $\mathbf{r} = (r_0, r_1, \dots, r_n) \in \mathbb{R}_+^{n+1}$ . The vector  $\mathbf{r}$  is the decision-maker’s subjectively perceived vector of “fair reference payoffs”; that is, a measure of how deserving she thinks each of the individuals is (including herself). The task of the axioms will be twofold. First, they will constrain which rankings  $\succeq_{\mathbf{r}}$  are permissible for a given  $\mathbf{r}$  (this axiomatises outcome-based fairness concerns); second, they will determine how the ranking changes when  $\mathbf{r}$  is altered

---

particular intuition is in fact accommodated by Axiom RH in section 4, which says that preferences are invariant with respect to equiproportional changes in deservingness, so that only relative reference payoffs matter.

(this axiomatises the sensitivity of fairness preferences to non-outcome factors that alter the deservingness of the individuals in  $N_0$ ). We shall not model how  $\mathbf{r}$  itself varies with any of the particular non-outcome factors mentioned above, although I do give one example (case 5 below).

We only consider preference structures that possess the following regularity properties<sup>6</sup>:

(1) *Completeness of preference relations.* For all  $\mathbf{r} \in \mathbb{R}_+^{n+1}$ ,  $\succsim_{\mathbf{r}}$  is complete:

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}, \mathbf{x} \succsim_{\mathbf{r}} \mathbf{y} \text{ or } \mathbf{y} \succsim_{\mathbf{r}} \mathbf{x}.$$

(2) *Transitivity of preference relations.* For all  $\mathbf{r} \in \mathbb{R}_+^{n+1}$ ,  $\succsim_{\mathbf{r}}$  is transitive:

$$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}_+^{n+1}, [\mathbf{x} \succsim_{\mathbf{r}} \mathbf{y} \text{ and } \mathbf{y} \succsim_{\mathbf{r}} \mathbf{z}] \rightarrow \mathbf{x} \succsim_{\mathbf{r}} \mathbf{z}.$$

(3) *Continuity of preference relations.* For all  $\mathbf{r}, \mathbf{x} \in \mathbb{R}_+^{n+1}$ , the sets  $\{\mathbf{y} \in \mathbb{R}_+^{n+1} \mid \mathbf{y} \succsim_{\mathbf{r}} \mathbf{x}\}$  and  $\{\mathbf{z} \in \mathbb{R}_+^{n+1} \mid \mathbf{x} \succsim_{\mathbf{r}} \mathbf{z}\}$  are closed.

(4) *Continuity of the preference structure.* For all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$ , the set  $\{\mathbf{r} \in \mathbb{R}_+^{n+1} \mid \mathbf{x} \succsim_{\mathbf{r}} \mathbf{y}\}$  is closed.

### 3. OUTCOME-BASED FAIRNESS PREFERENCES

As explained above, the outcome-based theories can be treated as the limiting case of reference-independence. In this section, we confine our attention to this limit case. Formally, we impose the following axiom:

**Axiom 1.** *Reference-independence (I).*

For all reference points  $\mathbf{r}, \mathbf{r}' \in \mathbb{R}_+^{n+1}$  and for all payoff distributions  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$ :

$$\mathbf{x} \succsim_{\mathbf{r}} \mathbf{y} \iff \mathbf{x} \succsim_{\mathbf{r}'} \mathbf{y}$$

This requires the decision-maker not to care about how much people deserve.

Now if she were completely self-interested, her preferences would be invariant to all transformations of the payoff vector that leave her own payoff unchanged.<sup>7</sup> We know, however, that people are not completely self-interested. How far do we need to depart from the conventional self-interested model to capture the empirically observed behaviour? It is reasonable to think that in pairwise

<sup>6</sup>These properties are numbered C1, C2, C5 and C6 in Munro & Sugden (2003).

<sup>7</sup>Consider the following axiom of reference-independent selfishness: for all  $\mathbf{x}, \mathbf{y}, \mathbf{r} \in \mathbb{R}_+^{n+1}$ ,  $x_0 \geq y_0 \iff \mathbf{x} \succsim_{\mathbf{r}} \mathbf{y}$ . This axiom entails all the invariance axioms considered here.

comparisons, a preference for one payoff distribution over the other should at least reflect the situations of those that are treated *differently* in the two distributions. If the preference reflects *only* those differences, it satisfies an axiom of separability:<sup>8</sup>

**Axiom 2.** *Strong Separability of Unaffected Individuals (S).*

For any reference point  $\mathbf{r} \in \mathbb{R}_+^{n+1}$ , for any set of individuals  $I \subseteq N_0$ , and for all payoff distributions  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  such that  $\mathbf{x}_I = \mathbf{y}_I$ :

$$\forall \mathbf{z}_I \in \mathbb{R}_+^{|I|} : [\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff (\mathbf{z}_I, \mathbf{x}_{\sim I}) \succeq_{\mathbf{r}} (\mathbf{z}_I, \mathbf{y}_{\sim I})]$$

Axiom S is very restrictive. We should note, however, that it is a *relaxation* of the conventional (selfish) model. If it can generate the kinds of utility functions calibrated in the experimental literature, moreover, its restrictiveness is a strength, as it captures behaviour in the most parsimonious way possible. I illustrate below that many important studies of outcome-based distributive preferences do indeed find that preferences are by and large compatible with S.

A possible objection to S is that even though the absolute level of unaffected individuals may be irrelevant, people often seem to care about their own *relative* standing. Experimental subjects are often found, for example, to have higher concern for those whose payoffs are lower than theirs, than for those who are ahead of them (Fehr & Schmidt 1999). Consider the following definition:

**Definition 1.** For a subset  $H \subseteq N$  of the set of individuals  $N$ , the set of payoff distributions that are *rank-maintaining relative to the decision-maker* is defined by

$$\mathbb{R}_H \equiv \{ \mathbf{x} \in \mathbb{R}_+^{n+1} \mid x_i \leq x_0 \iff i \in H \}$$

A simple weakening of S (and further relaxation of selfishness) imposes invariance to *rank-maintaining*<sup>9</sup> payoff levels of unaffected individuals:

<sup>8</sup>This rules out, in particular, utility functions that use payoff *differences* as irreducible arguments, such as those axiomatised by Neilson (2002, 2006). The separability axiom therefore disallows this kind of “building in” distributive concerns into the domain of the utility function.

<sup>9</sup>By choosing other notions of rank-maintaining payoff distributions, we could generate even weaker separability axioms. One could for example confine separability to hold within payoff sets that are invariant to the individuals’ ranks within the entire group (not just vis-à-vis the decision-maker). Such notions of rank-dependence have been used in non-expected utility theory, as mentioned in section 5. These extensions are beyond the scope of the present paper, but are interesting topics for further research.

**Axiom 3.** *Weak Separability of Unaffected Individuals (WS).*

For all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$ , for any sets of individuals  $H \subseteq N, I \subseteq N_0$ , and for all payoff distributions  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}_H$  such that  $\mathbf{x}_I = \mathbf{y}_I, (\mathbf{z}_I, \mathbf{x}_{\sim I}) \in \mathbb{R}_H$  and  $(\mathbf{z}_I, \mathbf{y}_{\sim I}) \in \mathbb{R}_H$ :

$$\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff (\mathbf{z}_I, \mathbf{x}_{\sim I}) \succeq_{\mathbf{r}} (\mathbf{z}_I, \mathbf{y}_{\sim I})$$

As I illustrate below, nearly all the common outcome-based utility functions used in the empirical literature satisfy one or both of these axioms.

In the situation we are considering—where the decision-maker does not care about deservingness—there is no reason for a decision-maker not to treat all other individuals equally. Thus it is natural to impose an axiom of *neutrality*. From a social choice perspective, of course, neutrality would be an ethically appealing requirement as well. In this positive theory we can justify it as rationally required when there are no relevant differences between the various “others”—as is the case in most experimental situations, where researchers typically take care to remove any knowledge of other participants’ individual characteristics.<sup>10</sup>

**Axiom 4.** *Payoff Neutrality (N).*

Let  $p : N \xrightarrow{\text{onto}} N$  be a permutation of the individuals other than the decision-maker. For any such  $p$ , and for all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$  and all payoff distributions  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$ :

$$\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff (x_0, x_{p(1)}, \dots, x_{p(n)}) \succeq_{\mathbf{r}} (y_0, y_{p(1)}, \dots, y_{p(n)})$$

There is much evidence that in evaluating payoff vectors, people mostly concern themselves with *relative* payoffs. Indeed there is a venerable tradition in economics exploring this tendency, which has recently received new attention as “catching-up-with-the-Joneses preferences” (Ok & Koçkesen 2000). From a positive perspective, we find that behaviour in experimental games is remarkably robust to equiproportional increases in the distributive outcomes, even for experimental research carried out in poor countries where the stakes can be equivalent to several weeks’ worth of

<sup>10</sup>Of course, when those characteristics are known, decision-makers can and often do discriminate between the other people involved. If the non-neutrality is *systematic*, however, it can be captured as reference-dependence. Sexist preferences (studied by Andreoni & Petrie (2004)), for example, can be captured by differentiating the values in the reference vector  $\mathbf{r}$  for men and for women, and relax the axioms of reference-independence, as we do in the following section.

income (Camerer 2003, Chapter 2, provides a survey). Accordingly, we restrict a decision-maker's concern with distributive outcomes to the *relative* payoffs of others:

**Axiom 5.** *Payoff Homotheticity (H).*

For all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$  and for all payoff distributions  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$ :

$$\forall \lambda > 0 : \mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff \lambda \mathbf{x} \succeq_{\mathbf{r}} \lambda \mathbf{y}$$

This axiom says that if one distribution is ranked above another, equiproportional increases or decreases of those distributions will be ranked in the same order as long as they are evaluated from the perspective of the same reference point.

Finally, it is conventionally assumed that more payoff is better. When people care about the payoffs of others, however, a higher payoff to oneself need not be better if it comes at the cost of much lower payoffs to others. A weaker version of increasingness, which seems reasonable both on positive and normative grounds, is to require that when distributive concerns are *not* at stake—when a decision-maker is choosing between two perfectly egalitarian allocations—more payoff is better. Formally:

**Axiom 6.** *Minimal Payoff Increasingness (MI).*

For all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$  and all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  such that  $\mathbf{x} = \{x_0, x_0, \dots, x_0\}$  and  $\mathbf{y} = \{y_0, y_0, \dots, y_0\}$ :

$$x_0 > y_0 \iff \mathbf{x} \succ_{\mathbf{r}} \mathbf{y}$$

Combining the axioms yields the following representation theorem:

**Theorem 1.** *If  $n \geq 2$ , then a preference structure  $\succeq$  satisfies axioms I, S, N, H, and MI if and only if it can be represented by a utility function of the form:*

$$U(\mathbf{x}) = \begin{cases} \text{sign}(\rho) \left[ (1 - n\alpha) x_0^\rho + \alpha \sum_{i=1}^n x_i^\rho \right] & \text{if } \rho \neq 0 \\ \text{or} \\ (1 - n\alpha) \ln x_0 + \alpha \sum_{i=1}^n \ln x_i \end{cases}$$

such that  $\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff U(\mathbf{x}) \geq U(\mathbf{y})$  for all payoff allocations  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  and for all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$ .

This family of distributive utility functions is the constant elasticity of substitution (CES) functional form, defined over payoff distributions. The parameters have a natural interpretation.  $\alpha$  is an altruism weight, with  $\alpha / (1 - n\alpha)$  measuring the marginal rate of substitution between the decision-maker's and another person's payoff at allocations where they get the same amount. If  $\alpha = 0$ , the utility function reduces to the special case of pure self-interest. Given some altruism, however,  $\rho$  is a measure of *inequality-aversion*. Provided  $\rho < 1$ , the decision-maker will place a higher value on a marginal dollar to another person the further ahead of him she is, and *vice-versa*. In the limit as  $\rho \rightarrow -\infty$ , the utility function comes to represent Leontief (or ‘‘Rawlsian’’) preferences, with the entire weight on the payoff of the worst-off person, no matter how small the differences are.

If we weaken the axiomatic basis by substituting WS for S, we generate a slightly more general CES structure:

**Theorem 2.** *If  $n \geq 2$ , then a preference structure  $\succeq$  satisfies axioms I, WS, N, H, and MI if and only if it can be represented by a utility function of the form:*

$$U(\mathbf{x}) = \begin{cases} \text{sign}(\rho) \left[ (1 - h\alpha - (n - h)\beta) x_0^\rho + \alpha \sum_{\{i \in N : x_i \leq x_0\}} x_i^\rho + \beta \sum_{\{i \in N : x_i > x_0\}} x_i^\rho \right] & \text{if } \rho \neq 0 \\ \text{or} \\ (1 - h\alpha - (n - h)\beta) \ln x_0 + \alpha \sum_{\{i \in N : x_i \leq x_0\}} \ln x_i + \beta \sum_{\{i \in N : x_i > x_0\}} \ln x_i \end{cases}$$

with  $h \equiv |\{i \in N : x_i \leq x_0\}|$

such that  $\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff U(\mathbf{x}) \geq U(\mathbf{y})$  for payoff allocations  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  and for all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$ .

This axiomatic basis unifies most of the functions proposed in the literature to rationalise the experimental evidence. I here survey the most important ones.

**Case 1.** *Andreoni and Miller's (2002) heterogeneous social utility function*

Andreoni & Miller (2002) calibrate a 2-person version of the utility function in theorem 1:

$$U(\mathbf{x}) = [(1 - \alpha)x_{self}^\rho + \alpha x_{other}^\rho]^{1/\rho}$$

where  $x_{self}$  is the payoff to oneself,  $x_{other}$  is the payoff to the other person, and  $\alpha$  is an altruism parameter.<sup>11</sup> They estimate the parameters and find perfect fits for three subgroups of agents, and good fits for three further subgroups that are “weak versions” of the three ideal types. The three types are selfish types, who have a value of  $\alpha$  close to zero and  $\rho$  close to 1; Leontief or egalitarian types, for whom  $\rho$  is strongly negative; and perfect-substitutes or utilitarian types, for whom  $\rho$  is close to one and  $\alpha$  close to .5.

**Case 2.** *Charness and Rabin’s (2002) 2-person “social utility function” without reciprocity*

Charness & Rabin (2002) present experimental data with which they calibrate a utility function of the form:<sup>12</sup>

$$U(\mathbf{x}) = \begin{cases} (1 - \alpha)x_{self} + \alpha x_{other} & \text{if } x_{other} \leq x_{self} \\ (1 - \beta)x_{self} + \beta x_{other} & \text{if } x_{other} > x_{self} \end{cases},$$

The use the label “social welfare preferences” for the parameter range  $1 \geq \alpha \geq \beta \geq 0$ , and “inequality-aversion” for the parameter range  $1 \geq \alpha \geq 0 \geq \beta \geq -1$ .

This utility function corresponds exactly to the functional form derived in theorem 2, with  $\rho = 1$  and  $n = 1$ .<sup>13</sup>

<sup>11</sup>Their exact formulation is  $U(\mathbf{x}) = (ax_s^\rho + (1 - a)x_o^\rho)^{1/\rho}$  where  $a$  is a “selfishness parameter.” For presentational and consistency reasons, I use  $\alpha$  as the altruism parameter. Andreoni and Miller use the conventional CES form, normalised by raising the function to the power  $1/\rho$  to be linearly homogeneous. In the absence of uncertainty (which they assume in their estimation) their normalisation does not affect the choice behaviour captured by the utility function in theorem 1. The normalisation used in the present theorems, however, permits a larger domain when the altruism weight is negative.

<sup>12</sup>Their notation uses  $\rho$  for the weight on the other person’s payoff when the decision-maker is ahead, and  $\sigma$  when she is behind.

<sup>13</sup>Note that their “social welfare preferences” can be approximated even by the functional form derived in theorem 1. Their parameter range ensures that the marginal rate of substitution is higher when the agent is ahead ( $\frac{\alpha}{1 - \alpha} \frac{x_{other}}{x_{self}}$ ) than when she is behind ( $\frac{\beta}{1 - \beta} \frac{x_{other}}{x_{self}}$ ). The same intuition—that people are more generous when they are ahead than when they are behind—is fully captured by the utility function in theorem 1, provided  $\rho < 1$ . The CES function therefore provides a smooth version of Charness and Rabin’s social welfare preferences with the same number of parameters

**Case 3.** *Fehr and Schmidt's (1999) many-person inequality-averse utility function*<sup>14</sup>

Fehr & Schmidt (1999) use the following utility function to capture inequality-averse behaviour:

$$U(\mathbf{x}) = x_0 + \alpha \sum_{i=1}^n \text{Min} [(x_i - x_0), 0] + \beta \sum_{i=1}^n \text{Max} [(x_i - x_0), 0]$$

which is just the functional form in theorem 2 with  $\rho = 1$ .

**Case 4.** *Charness and Rabin's (2002) "sophisticated social utility function" without reciprocity, smooth approximation*

In the appendix to their 2002 paper, Charness and Rabin also propose a more complex utility function than the 2-person piecewise linear utility function above. They assume that an agent's preferences can be represented by a weighted average of her own payoff, the total payoff of all individuals summed together, and the payoff of the worst-off individual. This utility function—sometimes not altogether felicitously named “Rawlsitarian” (Camerer 2003)—takes the following form:

$$U(\mathbf{x}) = (1 - \lambda) x_0 + \lambda \left[ \delta \min_{i \in N_0} x_i + (1 - \delta) \sum_{i \in N_0} x_i \right].$$

Again, the functional form is intended to parametrise an intermediate case between self-interest, social welfare/utilitarian preferences, and Leontief/“Rawlsian” preferences.  $\lambda$  denotes the weight on distributive concerns, within which  $\delta$  measures the weight on the Leontief preferences. The same trade-offs are captured by the functional form in theorem 1. The altruism parameter  $\alpha$  plays the same role in theorem 1 as  $\lambda$  does above; whereas a strongly negative  $\rho$  in the CES function corresponds to a  $\delta$  close to 1 above, and  $\rho$  approaching 1 corresponds to a  $\delta$  close to zero. Thus

---

(two), and in fact the smooth form has been shown to fit their data much more precisely than their own functional form (Sandbu 2007).

<sup>14</sup>Theorem 2 provides a different axiomatic basis for the Fehr-Schmidt inequality-aversion model than the axiomatisation proposed by Neilson (2002, 2006). Instead of a separability axiom, Neilson uses an axiom of “self-referent separability,” which imposes separability in payoff *differences* between the decision-maker and the other players. This approach allows him to derive a class of inequality-averse distributive utility functions of which the Fehr-Schmidt model is a linear special case.

the CES function is a smooth alternative to Charness and Rabin’s linear/Leontief form, which in experiments with continuous choice sets would allow for interior solutions without reducing the degrees of freedom.

The most important outcome-based theory that is *not* covered by one of the two theorems is Bolton and Ockenfels’ (2000) theory of “Equity, Reciprocity, and Competition” (ERC). They do not fully specify a functional form, but their generic function distinguishes itself from the Fehr-Schmidt approach in assuming that people care only about inequality as far as it concerns themselves. They postulate a “motivation function”  $v = v_0(y_0, \sigma_0)$  where the arguments are the agent’s payoff level  $y_0$  and her share in *aggregate* payoff:

$$\sigma_0 = \begin{cases} y_0 / \sum_{i=0}^n y_i & \text{if } \sum_{i=0}^n y_i > 0 \\ 1/n & \text{if } \sum_{i=0}^n y_i = 0 \end{cases}$$

The conditions they impose on the motivation function produce inequality-aversion. In particular, the maximand  $v$  is increasing in  $\sigma_0$  when  $\sigma_0 < 1/n$ , and decreasing in  $\sigma_0$  for  $\sigma_0 > 1/n$ . This assumption violates both of the separability axioms used in this section. (But as case 3 shows, the intuition of inequality-aversion can be modelled as a special case of the specification in theorem 2.)

It turns out, then, that most extant specifications of distributive preferences can be unified in a family of functions characterised by a simple set of axioms. The axioms have natural interpretations in terms of other-regarding attitudes. The relevance for social choice theory is twofold. Some of the axioms may be normatively attractive at the level of social aggregation of individual preferences, or may have normative appeal to the extent that they are held by individuals.<sup>15</sup> Even if they are not, however, they can be employed in empirical social evaluation exercises to separate those components of preferences that are motivated by fairness concerns from those that are not, in order to treat them differentially in social aggregation.

---

<sup>15</sup>Indeed equivalents of several of these axioms have been applied in social choice theory—see section 5—and the present contribution may be seen as validating those axioms in the sense that they seem to govern people’s actual views of fairness.

## 4. REFERENCE-DEPENDENT FAIRNESS PREFERENCES

We now proceed to axiomatising non-outcome-based models of distributive preferences. Recall that we model the individual as capturing non-outcome factors in an overall judgment of what each individual deserves (the reference allocation). This approach makes a conceptual distinction between what people deserve in the decision-maker's mind, and what they receive at her hands. In a Dictator game situation, for instance, she may consider that both she (as the Dictator) and the recipient deserve the same, yet deviate from an equal split to her own advantage. If individuals' perceptions of what people deserve could be extrapolated from their choices or expressed preferences over states of affairs, that would be an appealing input into theories of social aggregation—more appealing, perhaps, than “unfiltered” preferences. Such extrapolations, however, require mathematical foundations for the functional forms to be calibrated from observed behaviour. In this section I propose an axiomatisation of how changes in the fair reference allocation affect preferences over actual payoffs—the axioms considered here are thus imposed on the functional relationship between given reference points and orderings of payoff vectors.<sup>16</sup> Positive results give less guidance to choosing axioms in this case than they did in the previous section. This is because the empirical research has mostly focused on one non-outcome factor at a time (*e.g.* reciprocity), rather than on how non-outcome factors in general affect preferences. Instead I suggest that the axioms applied in the outcome-based case are naturally extended to apply to the preference structure as a whole. Each of the axioms discussed in section 3 impose an invariance requirement on the preference *relations* over different payoff distributions, and I shall argue that in each case, there are reasons to impose the same kind of invariance also on the effect of the reference allocation on the preference relations. A decision-maker who obeys the axioms for reference-independence can reasonably be expected, on a presumption of consistency, to obey their extension to reference-dependence.

---

<sup>16</sup>It bears emphasising that this is a modest approach. We are not explaining or axiomatising how the reference vector itself summarises non-outcome or contextual factors. That it is beyond the scope of this exercise, which aims merely at showing how generic contextual influences could be accommodated within the same formal structure that has been used for outcome-based theories. It suggests, to be precise, that if contextual factors can be summarised as a reference vector, they can be incorporated into utility functions defined over payoff allocations as described below, and conversely, they can be extrapolated from observed choice behaviour in appropriately designed experiments.

Consider a pairwise ranking between two payoff distributions  $\mathbf{x}$  and  $\mathbf{y}$  that give the same payoff to some individuals, and suppose the reference point changes—some people become more or less deserving. If the only people whose deservingness changes are those who get the same payoff in both allocations, how is the preference ranking to be affected? The reasoning behind the separability axiom in the previous section suggests that it should not change. There, we argued that when the *payoffs* of unaffected individuals change, the ranking is unaffected. For an individual who obeys that axiom (as the evidence cited above suggests people do), it would seem inconsistent if, in contrast, she let her ranking be affected when *reference* payoffs of unaffected individuals change. Formally, we can express this intuition as a separability axiom that resembles axiom S, but now applied to the changes in the reference point:

**Axiom 7.** *Strong Independence of Irrelevant Reference Payoffs (IIR).*

*For any set of individuals  $I \subseteq N_0$ , for all payoff allocations  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  such that  $\mathbf{x}_I = \mathbf{y}_I$  and for all reference points  $\mathbf{r}, \mathbf{r}' \in \mathbb{R}_+^{n+1}$  such that  $\mathbf{r}_{-I} = \mathbf{r}'_{-I}$ :*

$$\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff \mathbf{x} \succeq_{\mathbf{r}'} \mathbf{y}$$

While IIR is restrictive, it constitutes a relaxation of the conventional and near-universal assumption of reference-independence, just like S constituted a relaxation of the conventional assumption of selfishness.<sup>17</sup> It is desirable in an axiomatic exercise like this one to investigate how much we can capture with only minimal departures from the standard model. Still, weaker separability axioms—implying a further departure from reference-independence—could be considered. Given

<sup>17</sup>What reason could a decision-maker have for violating IIR? The change in reference payoffs means that some individuals will now fall short or exceed their deserved amount more or less than before. But the choice between  $\mathbf{x}$  and  $\mathbf{y}$  can do nothing to respond to that. So if the decision-maker changes her preference, it must mean she tries to “compensate” for how some individuals are moved closer to or further away from what they deserve (because of the change from  $\mathbf{r}$  to  $\mathbf{r}'$ ) by moving *other* individuals closer to or further away from what *they* deserve (which has not changed). But this would allow for bizarre choices. Suppose I can allocate 40 dollars between four individuals in one of two ways: Either 10 dollars to each person, or \$15 to myself, \$5 to person A, and \$10 each to person B and person C. Originally I judge us as each deserving \$10, but I prefer the allocation more favourable to myself because getting five extra dollars is more valuable to me than giving everyone what they deserve. Now suppose person B acts kindly to me in some (non-pecuniary) way and person C acts unkindly to me. As in the example mentioned in section 2.1, I now judge B more deserving and C less deserving than before. What IIR rules out is that I shift my preference to the egalitarian allocation in order, as it were, to mitigate B and C’s not getting what *they* deserve by bringing myself and A closer to what *we* deserve. But this seems to be a misplaced response to the fact of B and C not getting what they deserve—my choice only addresses the relative deservingness of A and myself, which I earlier considered less valuable than an extra five dollars to myself.

the special place of the decision-maker's own payoffs in her preferences, a plausible weaker axiom allows changes in the reference point for the decision-maker,  $r_0$ , to matter even if the decision-maker's payoff is the same in two allocations:

**Axiom 8.** *Weak Independence of Irrelevant Reference Payoffs (WIIR).*

For all sets of individuals  $I \subseteq N$ , for all payoff allocations  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  such that  $\mathbf{x}_I = \mathbf{y}_I$  and for all reference points  $\mathbf{r}, \mathbf{r}' \in \mathbb{R}_+^{n+1}$  such that  $\mathbf{r}_{-I} = \mathbf{r}'_{-I}$ :

$$\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff \mathbf{x} \succeq_{\mathbf{r}'} \mathbf{y}$$

Beyond WIIR, one could consider even weaker independence axioms defining narrower invariance classes; we leave this matter for further research.

Once the reference payoffs are allowed to matter, the motivation for the neutrality axiom N must be reconsidered. In this more general case, there is no reason why a decision-maker should be neutral with respect to a reshuffling of payoffs. In particular, she presumably prefers a reshuffling that gives the highest payoffs to the most deserving individuals. The natural modification of the neutrality axiom is to impose invariance if the *pairings* of reference payoffs and actual payoffs are shuffled among the other individuals. If the reference point captures all the differences in how much people deserve, there is no reason for a rational decision-maker to treat individuals differently once their reference claims have been accounted for. We formalise this intuition as:

**Axiom 9.** *Reference Point Neutrality (RN).*

Let  $p : N \xrightarrow{\text{onto}} N$  be a permutation of the individuals other than the decision-maker. For any such  $p$ , for all payoff allocations  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  and all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$ :

$$\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff (x_0, x_{p(1)}, \dots, x_{p(n)}) \succeq_{(r_0, r_{p(1)}, \dots, r_{p(n)})} (y_0, y_{p(1)}, \dots, y_{p(n)})$$

In words, a preference order is preserved if both reference payoffs and actual payoffs are shuffled among the other individuals, with the pairing of actual payoffs and reference claims remaining unchanged.

The homotheticity axiom on preference relations introduced in the previous section extends naturally to a homotheticity requirement for the preference structure as a whole. If what matters to a decision-maker's other-regarding preferences are the ratios of *actual* payoffs, as homotheticity entails, then it stands to reason that it is also the ratios of how much she thinks each individual *deserves* that influence her preferences. We capture this intuition in the following axiom:

**Axiom 10.** *Reference Point Homotheticity (RH).*

For all payoff distributions  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  and for all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$ :

$$\forall \lambda > 0 : \mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff \mathbf{x} \succeq_{\lambda \mathbf{r}} \mathbf{y}$$

In essence, these three axioms (IIR or WIIR, RN, and RH) do to reference-independence what the invariance axioms in the previous section (S or WS, N, and H) do to selfishness. The latter work at the level of the individual preference *relation*, while the former apply to the preference *structure* as a whole. Both sets of axioms define a similar invariance class to weaken the conventionally assumed extreme invariance with respect to what other people deserve or what they actually receive, respectively. Put differently, all the axioms are entailed by selfishness or by reference-independence.<sup>18</sup>

We complete this parallelism by introducing an increasingness axiom for the preference structure. It is a very natural one—when one person's fair reference payoff increases, that (weakly) tilts the preference order in his favour, other things being equal:

**Axiom 11.** *Reference-Point Increasingness (RPI).*

For any individual  $i \in N_0$ , for all reference points  $\mathbf{r}, \mathbf{r}' \in \mathbb{R}_+^{n+1}$  such that  $r'_i > r_i$  and  $\mathbf{r}'_{\sim i} = \mathbf{r}_{\sim i}$ , and for all payoff allocations  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  such that  $\mathbf{x} \sim_{\mathbf{r}} \mathbf{y}$ :

$$x_i \geq y_i \iff \mathbf{x} \succeq_{\mathbf{r}'} \mathbf{y}$$

We can now discuss a representation theorem that uses the weak version of independence of irrelevant reference points:

<sup>18</sup>Axiom RN is entailed by the conjunction of neutrality (axiom N) and reference-independence, and *a fortiori* by the conjunction of selfishness and reference-independence.

**Theorem 3.** *If  $n \geq 2$ , a preference structure  $\succeq$  satisfies axioms S, WIIR, RN, H, RH, MI and RPI if and only if for all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$  the corresponding preference relation  $\succeq_{\mathbf{r}}$  can be represented by a utility function of the form:*

$$U_{\mathbf{r}}(\mathbf{x}) = \begin{cases} \text{sign}(\rho) \left[ \left( 1 - \sum_{i=1}^n \alpha_i(\mathbf{r}) \right) x_0^\rho + \sum_{i=1}^n \alpha_i(\mathbf{r}) x_i^\rho \right] & \text{if } \rho \neq 0 \\ \text{or} \\ \left( 1 - \sum_{i=1}^n \alpha_i(\mathbf{r}) \right) \ln x_0 + \sum_{i=1}^n \alpha_i(\mathbf{r}) \ln x_i \end{cases}$$

where

$$\alpha_i(\mathbf{r}) \equiv \frac{\phi(r_i/r_0)}{1 + \sum_{j=1}^n \phi(r_j/r_0)},$$

and either  $\max_{z \geq 0} \phi(z) < -1/n, \phi' \leq 0$

or  $\min_{z \geq 0} \phi(z) > -1/n, \phi' \geq 0,$

such that  $\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff U_{\mathbf{r}}(\mathbf{x}) \geq U_{\mathbf{r}}(\mathbf{y})$  for all payoff allocations  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$ .

It is simple to see that if the fair reference payoffs  $\mathbf{r}$  are constant (or the functions  $\alpha_i(\cdot)$  are degenerate), the weights  $\alpha_i$  are constant, and so this functional form reduces to the one we derived in theorem 1. In other words, the outcome-based models constitute the limiting case of reference-dependence in which what people deserve (according to the decision-maker) is unaffected by any other factors—that is, reference-*independence*. Beyond this special case, the functional form in theorem 3 can capture different types of non-outcome-based theories of fairness preferences, by suitably specifying how the  $\mathbf{r}$ -vector is determined. One may, for example, let it depend on the “kindness” or “fairness” of previous actions, by making  $\mathbf{r}$  an appropriate function of previous actions and the game form. This would capture the intuitions behind the reciprocity theories of Rabin (1993) and Falk & Fischbacher (2000) in a formal framework that is at the same time compatible with the outcome-based theories surveyed in section 3.

An interesting example is found in Cox *et al.* (2007), who use a utility function similar to, but less general, than the one derived in theorem 1. In their “tractable model of reciprocity and fairness,” they propose that the MRS between own and others’ payoffs depends on reciprocity and

“status.” The present work can be seen as providing axiomatic foundations for Cox *et al.*’s attempt at general yet tractable fairness models; their specific model can easily be modified to fit as a special case of the forms derived in this paper.

The last case to be considered is an empirical application of the functional form of theorem 3:

**Case 5.** Sandbu’s (2007) set-dependent fairness preferences

A particularly simple functional form for  $\phi$  is  $\phi(r_i/r_0) = a + c(r_i/r_0)$ , which yields reference-dependent utility functions of the form:

$$U_{\mathbf{r}}(\mathbf{x}) = \text{sign}(\rho) \left[ \left( 1 - \sum_{i=1}^n \alpha_i(\mathbf{r}) \right) x_0^\rho + \sum_{i=1}^n \alpha_i(\mathbf{r}) x_i^\rho \right]$$

with

$$\alpha_i(\mathbf{r}) \equiv \frac{a + c(r_i/r_0)}{1 + na + c \sum_{j=1}^n (r_j/r_0)}$$

$$a > -1/n, c \geq 0$$

(we omit the Cobb-Douglas case). A two-person version of this functional form is investigated by Sandbu (2007), who multiplies by  $\left[ (1 + na) + c \sum_{j=1}^n (r_j/r_0) \right]$  to get the normalised utility function:

$$U_{\mathbf{r}}(\mathbf{x}) = \text{sign}(\rho) \left[ x_{self}^\rho + \left( a + c \frac{r_{other}}{r_{self}} \right) x_{other}^\rho \right]$$

The reference point in this study is the most egalitarian allocation in the subset of achievable allocations that are efficient and do not put the decision-maker behind the other player. The intuition is that the *availability* of fair outcomes makes fairness more salient and therefore increases the weight on other people’s payoffs. Experiments confirm that when the reference point becomes less egalitarian, subjects do tend to behave more selfishly when choosing even among the allocations that remain available. This functional form lends itself readily to interpretation. The weight the decision-maker puts on another person’s payoff when he deserves the same as her ( $r_i = r_0$ ) is  $a + c$ . This magnitude may therefore be thought of as

measuring *bias*: If  $a + c = 1$ , then the decision-maker prefers to share equally with those she considers equally deserving as her, whereas if  $a + c < 1$ , she is biased in her own favour. If  $r_i = 0$ , that is, if person  $i$  does not deserve any payoff at all,  $a$  alone measures the weight on  $i$ 's payoff and can therefore be interpreted as the degree of pure benevolence (or malevolence if  $a < 0$ ). Finally,  $c$  measures the sensitivity to changes in the reference claim ratio. Note that if  $r_0 \rightarrow 0$ ,  $\frac{r_i}{r_0} \rightarrow \infty$ , which means that as long as  $c \neq 0$ , all the weight will be put on the payoff of other people when the decision-maker considers that she herself does not deserve anything.

If we strengthen the axiomatic basis by substituting IIR for WIIR, we can narrow the functional form down further:

**Theorem 4.** *If  $n \geq 2$ , a preference structure  $\succeq$  satisfies axioms S, IIR, RN, H, RH, MI and PRI if and only if for all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$  the corresponding preference relation  $\succeq_{\mathbf{r}}$  can be represented by a utility function of the form:*

$$U_{\mathbf{r}}(\mathbf{x}) = \begin{cases} \text{sign}(\rho) \left[ \left(1 - \sum_{i=1}^n \alpha_i(\mathbf{r})\right) x_0^\rho + \sum_{i=1}^n \alpha_i(\mathbf{r}) x_i^\rho \right] & \text{if } \rho \neq 0 \\ \text{or} \\ \left(1 - \sum_{i=1}^n \alpha_i(\mathbf{r})\right) \ln x_0 + \sum_{i=1}^n \alpha_i(\mathbf{r}) \ln x_i \end{cases}$$

where

$$\alpha_i(\mathbf{r}) \equiv \frac{a (r_i/r_0)^\gamma}{1 + \sum_{j=1}^n a (r_j/r_0)^\gamma},$$

$$a, \gamma \geq 0$$

such that  $\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff U_{\mathbf{r}}(\mathbf{x}) \geq U_{\mathbf{r}}(\mathbf{y})$  for all payoff allocations  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$ .

For the sake of completeness, we conclude this section with the representation theorem for reference-dependence with weak separability of unaffected individuals:

**Theorem 5.** *If  $n \geq 2$ , a preference structure  $\succeq$  satisfies axioms WS, WIIR, RN, H, RH, MI and RPI if and only if for all reference points  $\mathbf{r} \in \mathbb{R}_+^{n+1}$  the corresponding preference relation  $\succeq_{\mathbf{r}}$  can*

be represented by a utility function of the form:

$$U(\mathbf{x}) = \left\{ \begin{array}{l} \text{sign}(\rho) \left[ \begin{array}{l} \left( 1 - \sum_{x_i \neq 0 \leq x_0} \alpha_i(\mathbf{r}) - \sum_{x_i > x_0} \beta_i(\mathbf{r}) \right) x_0^\rho \\ + \sum_{\{i \in N : x_i \leq x_0\}} \alpha_i(\mathbf{r}) x_i^\rho + \sum_{\{i \in N : x_i > x_0\}} \beta_i(\mathbf{r}) x_i^\rho \end{array} \right] \\ \text{or} \\ \left( 1 - \sum_{x_i \neq 0 \leq x_0} \alpha_i(\mathbf{r}) - \sum_{x_i > x_0} \beta_i(\mathbf{r}) \right) \ln x_0 \\ + \sum_{\{i \in N : x_i \leq x_0\}} \alpha_i(\mathbf{r}) \ln x_i + \sum_{\{i \in N : x_i > x_0\}} \beta_i(\mathbf{r}) \ln x_i \end{array} \right. \quad \text{if } \rho \neq 0$$

where

$$\begin{aligned} \alpha_i(\mathbf{r}) &\equiv \frac{\phi_a(r_i/r_0)}{1 + \Phi(\mathbf{r})}, \\ \beta_i(\mathbf{r}) &\equiv \frac{\phi_b(r_i/r_0)}{1 + \Phi(\mathbf{r})}, \\ \Phi(\mathbf{r}) &\equiv \sum_{\{i \in N : x_i \leq x_0\}} \phi_a(r_i/r_0) + \sum_{\{i \in N : x_i > x_0\}} \phi_b(r_i/r_0), \end{aligned}$$

and either  $\max_{z \geq 0} \phi_{a,b}(z) < -1/n, \phi'_{a,b} \leq 0$

or  $\min_{z \geq 0} \phi_{a,b}(z) > -1/n, \phi'_{a,b} \geq 0$ ,

such that  $\mathbf{x} \succeq_{\mathbf{r}} \mathbf{y} \iff U_{\mathbf{r}}(\mathbf{x}) \geq U_{\mathbf{r}}(\mathbf{y})$  for all payoff allocations  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$ .

If the preference structure also satisfies axiom IIR, we have

$$\begin{aligned} \alpha_i(\mathbf{r}) &\equiv \frac{a(r_i/r_0)^\gamma}{1 + \Phi(\mathbf{r})}, \\ \beta_i(\mathbf{r}) &\equiv \frac{b(r_i/r_0)^\gamma}{1 + \Phi(\mathbf{r})}, \\ \Phi(\mathbf{r}) &\equiv \sum_{\{i \in N : x_i \leq x_0\}} a(r_i/r_0)^\gamma + \sum_{\{i \in N : x_i > x_0\}} b(r_i/r_0)^\gamma, \\ a, b, \gamma &\geq 0. \end{aligned}$$

## 5. RELATED LITERATURE

**5.1. Axiomatic approaches to fairness.** There exist a few other recent axiomatisations of individual fairness-motivated preferences. They include Neilson (2002, 2006), who provides a foundation for the Fehr-Schmidt inequality aversion model; Ok & Koçkesen (2000), who present an axiomatic derivation of the relative income hypothesis; and Karni & Safra (2002), who axiomatise a model of fairness judgments over lotteries for an indivisible good. All of these studies have chosen to investigate a specific model of fairness-motivated preferences, in contrast with the present approach, which has examined the underlying similarities between the different theories that have emerged from the experimental literature.

Individual preferences over payoff allocations bear an obvious affinity to social choice theory. Whereas the present approach derives individual evaluation functions defined over the payoffs to several individuals, social choice theory derives social choice functions or social evaluation functions defined over the utility indexes or preferences of several individuals. Up to a certain point, the two approaches are isomorphic, although the different interpretations of the axioms necessitate separate justification and may lead to other normative conclusions. Most closely related to the present paper is the analysis of Blackorby & Donaldson (1982). They prove a representation theorem that closely resembles Theorem 1 in this paper—so the present results can be taken to mean that empirical behaviour supports their axioms—but their result is slightly less general and the proof is different from the one offered here. Given certain conditions, they prove that a social evaluation function must be a global means of order  $t$ , with  $t > 0$  if negative utilities are admitted. In addition to technical requirements, they require the social evaluation function to satisfy “minimal increasingness” in individual utilities, “complete strict separability” in individual utility numbers (equivalent to our strong separability), and that individual utility numbers satisfy ratio-scale full comparability. The last axiom entails the homotheticity of the social evaluation function, which we in contrast impose directly as an axiom on the distributive utility function. There is an important interpretive difference. Ratio-scale comparability is justified on the basis of how individual utilities are thought to be comparable; it is an axiom on the informational content of the utility measure. In

the present approach, however, the domain of the distributive utility function is the space of monetary payoffs, which are evidently fully comparable. The present homotheticity axiom is therefore not a restriction on the informational content of the domain, but on ways in which that information is used, *viz.* a requirement that only ratio-scale comparable information should matter for the preferences. A further difference is that the special place of the decision-maker in her own preferences has no corresponding rationale in social choice theory, and so nor do the various modifications of the axioms that we explore as a consequence.

The social choice research program has given less attention to the possibility of reference-dependence in social choice.<sup>19</sup> There have, however, been some attempts to incorporate *procedural* concerns—which is part of what the reference point in the present model might capture—in particular by modelling the assignment of *rights* as an object of preferences over procedures (see Pattanaik & Suzumura (1994) and Suzumura (1999)). Reference-dependence in the form of set-dependence (case 5), moreover, has been analysed in axiomatic bargaining theory in the Kalai-Smorodinsky solution (Kalai & Smorodinsky 1975), which is sensitive to the northwest and the southeast corners of (the positive orthant of) the Pareto frontier.

**5.2. Non-expected utility theory.** Some contributions to the theory of choice under uncertainty are relevant to the analysis of reference-dependent fairness preferences. A number of alternatives to standard expected utility have been proposed, which have been shown to be variations of a common axiomatic basis (Chew & Epstein 1989). In one family of non-expected utility theory, the utilities of the different outcomes are weighted by a function that depends on the probability of the outcome and on the *rank* of the outcome in the agent's preferences. The rank-dependent theories are axiomatised with conditions of rank-dependent separability, which resemble our axiom WS. The main difference is that in individual choice under uncertainty there is no equivalent to the special role of the decision-maker in our model. The rank-dependence axioms in non-expected utility theory therefore introduce a sensitivity to the rank of each outcome in the *entire* set of possible outcomes. As suggested in section 3, an extension of the present work could consider

---

<sup>19</sup>Although see Plott (1973) for an investigation of path-*independence*, and Sen (1997) for an exploration of some formal properties of menu-dependent preferences and choice functions.

more rank-dependence in the separability axioms (*i.e.* narrower invariance classes). For these extensions, the axioms used in non-expected utility theory may be useful models.

**5.3. Reference-dependent utility in behavioural consumer theory.** Falk & Fischbacher (2000) and Cox *et al.* (2007) are the first studies of which I am aware that explicitly use the notion of a reference point for fairness in formal models of distributive preferences.<sup>20</sup> There are, however, earlier models of reference-dependent preferences in consumer theory, first developed by Tversky & Kahneman (1991). They have been further formalised, tested, and refined by Alistair Munro and Robert Sugden and their collaborators (Bateman *et al.* 1997, Munro & Sugden 2003). These models make preferences sensitive to a customary consumption bundle which acts as the reference point, and the studies show that reference-dependence can account for many of the systematic deviations from the predictions of neoclassical consumer theory, such as *status quo* bias and the endowment effect. The fact that reference-dependent theories seem successful as accounts of consumer choice suggests that reference-dependence is a deep-seated phenomenon that may characterise choice and preference generally.

## 6. CONCLUSION

The research programme on fairness-motivated preferences has produced a large body of knowledge of both observed behaviour and the structure of fairness motivations. This research makes possible a richer informational basis for social choice rules and social welfare functions than has typically been employed in social choice theory. I have proposed a decision-theoretic framework which can separate the fairness-motivated from the self-interested components of individual preferences. This separation can be fruitful for social choice theory by providing a way of “laundrying” preferences by removing private conflicts of interest from the domain of social preference aggregation. It could also provide the inputs to a social-choice exercise that made social choice rules depend on the axioms *governing* individuals’ preferences, rather than on just those preferences themselves.

<sup>20</sup>Brandts & Sola (2001) employ the notion in the motivation of an experimental study.

This paper conceptualises and formalises a general class of reference-dependent distributive preferences, and shows that most of the common outcome-based models can be seen as special limiting cases of the general model (as can the conventional self-interested, reference-independent, model). The axiomatic analysis characterises the underlying assumptions of this whole family of functions, and provides two sets of simple and plausible axioms—one imposing invariance on how preferences are sensitive to actual payoffs, and one imposing invariance on how preferences vary with fair reference payoffs—that are necessary and sufficient for preferences to be representable by the proposed utility functions. This work, then, lies in between two fields of enquiry in economics. At one end, it connects with social choice theory in the way described above. At the other end, it touches on behavioural and experimental economics, as the functional forms derived here need to be calibrated and tested in empirical research, while the axiomatisation provides mathematical foundations for the heuristic utility functions used in this literature. As such, the present article contributes to building closer connections between these two vast literatures.

#### APPENDIX A. APPENDIX: PROOFS

**A.1. Proof of Theorem 1.** It is clear from inspection that the functional form in theorem 1 satisfies the axioms. We here prove the necessity of the functional form. Note first that axiom I (reference-independence) implies that the preference structure is constant with respect to reference point changes, so there is a single preference relation  $\succeq$  ranking the elements of  $\mathbf{X}$ . Denote by  $U(\mathbf{x})$  the utility function that represents it.<sup>21</sup> By Debreu (1983, theorem 3) axiom S entails that it is additively separable:

$$(1) \quad U(\mathbf{x}) = T \left( \sum_{i=0}^n \nu_i(x_i) \right) \text{ where } T' > 0,$$

and axiom N entails  $\nu_i(\cdot) = \nu_a(\cdot), \forall i \in N$ .

We proceed by proving the following lemma:

---

<sup>21</sup>The existence of the utility function is entailed by the completeness, transitivity, and continuity of the preference relation, as proved by Debreu (1954).

**Lemma 1.**  $U(\mathbf{x})$  has a constant and identical direct elasticity of distribution between all of its arguments.

*Proof.* Consider two allocations  $\mathbf{x}$  and  $\mathbf{y}$  such that  $\mathbf{x} \sim \mathbf{y}$ . By axiom H, it must be the case that  $\lambda \mathbf{x} \sim \lambda \mathbf{y}$ , or

$$(2) \quad U(\lambda \mathbf{x}) = U(\lambda \mathbf{y}), \forall \lambda > 0.$$

Differentiating both sides of this expression with respect to  $\lambda$  and evaluating at  $\lambda = 1$  yields

$$(3) \quad \nu'_0(x_0)x_0 + \sum_{l=1}^n \nu'_a(x_l)x_l = \nu'_0(y_0)y_0 + \sum_{l=1}^n \nu'_a(y_l)y_l$$

for any points  $\mathbf{x}$  and  $\mathbf{y}$  on the same indifference surface. Now consider an individual  $j \in N$  and fix the incomes  $x_h$  of all other individuals  $h \in N \setminus \{j\}$ . Infinitesimally moving the point  $\mathbf{x}$  along the indifference curve in  $(x_0, x_j)$ -space while keeping  $\mathbf{y}$  unchanged yields

$$(4) \quad \left. \frac{d(\nu'_0(x_0)x_0)}{dx_0} + \frac{d(\nu'_a(x_j)x_j)}{dx_j} \frac{dx_j}{dx_0} \right|_{dU(\mathbf{x})=0, dx_h=0, \forall h \in N \setminus \{j\}} = 0$$

$$(5) \quad \nu''_0(x_0)x_0 + \nu'_0(x_0) - \frac{\nu'_0(x_0)}{\nu'_a(x_j)} (\nu''_a(x_j)x_j + \nu'_a(x_j)) = 0$$

so

$$(6) \quad \frac{\nu''_0(x_0)}{\nu'_0(x_0)}x_0 = \frac{\nu''_a(x_j)}{\nu'_a(x_j)}x_j = k$$

where the constancy follows from the fact that the equality holds at any point  $\mathbf{x}$  where  $\nu'_0(x_0) \neq 0$  and  $\nu'_a(x_j) \neq 0$ .

Choose any  $\{i, j\} \in N_0$  and let

$$\nu_i(\cdot) = \begin{cases} \nu_0(\cdot) & \text{if } i = 0 \\ \nu_a(\cdot) & \text{in } i \in N \end{cases}$$

Denote by  $\sigma_{ij}$  the direct elasticity of substitution between  $x_i$  and  $x_j$  and by  $MRS_{ij}$  the marginal rate of substitution between them:

$$\sigma_{ij} \equiv \frac{d \ln (x_j/x_i)}{d \ln |MRS_{ij}|}.$$

We have

$$(7) \quad \ln |MRS_{ij}| = \ln |\nu'_i(x_i)| - \ln |\nu'_j(x_j)|,$$

so

$$(8) \quad \sigma_{ij} = \frac{d(\ln x_j - \ln x_i)}{d(\ln |\nu'_i(x_i)| - \ln |\nu'_j(x_j)|)}$$

$$(9) \quad = \frac{\frac{dx_j}{x_j} - \frac{dx_i}{x_i}}{\frac{\nu''_i(x_i)}{\nu'_i(x_i)} dx_i - \frac{\nu''_j(x_j)}{\nu'_j(x_j)} dx_j}$$

$$(10) \quad = \frac{\frac{dx_j}{x_j} - \frac{dx_i}{x_i}}{\frac{k}{x_i} dx_i - \frac{k}{x_j} dx_j}$$

$$(11) \quad = -\frac{1}{k}$$

where line 10 follows from equation 6. So the direct elasticity of substitution is constant and identical between any two individuals' payoffs whenever it is defined.  $\square$

*Proof of the theorem.* McFadden (1963) generalises Arrow *et al.* (1961) and Uzawa (1962) and proves that an additively separable function has a constant and identical direct elasticity of substitution between all its arguments if and only if it can be written in the form:

$$(12) \quad f(\mathbf{x}) = \begin{cases} [\sum_{i=0}^n \delta_i x_i^\rho]^{1/\rho} & \text{if } \rho \neq 0 \\ \text{or} \\ \sum_{i=0}^n \delta_i \ln x_i \end{cases}$$

(or a monotonic transformation thereof). Together with this result, lemma 1 therefore entails that the utility function can be written in the form:

$$(13) \quad U(\mathbf{x}) = \begin{cases} T\left([\delta_0 x_0 + \delta_a \sum_{i=1}^n x_i^\rho]^{1/\rho}\right) & \text{if } \rho \neq 0 \\ \text{or} \\ T(\delta_0 \ln x_0 + \delta_a \sum_{i=1}^n \ln x_i) \end{cases}.$$

where  $T(\cdot)$  is any monotonic transformation. (We may verify by inspection that the derivation in lemma 1 is well-defined on  $\mathbb{R}_{++}^{n+1}$ , that is, on the entire domain of the preference relation except its boundaries.)

By axiom MI, we must have  $\delta_0 + n\delta_a > 0$ . We may therefore choose as a monotonic transformation the function  $T(z) \equiv \text{sign}(\rho) z^\rho (\delta_0 + n\delta_a)^{-1}$ , which yields theorem 1.<sup>22</sup>  $\square$

## A.2. Proof of Theorem 2.

*Proof of the theorem.* Sufficiency is obvious from inspection of the utility function in the theorem. To prove necessity, note first that axiom WS imposes separability *within* each subset of distributions

$$\mathbb{R}_H \equiv \left\{ \mathbf{x} \in \mathbb{R}_+^{n+1} \mid i \in H \subseteq N \longleftrightarrow x_i \leq x_0 \right\}.$$

Each  $\mathbb{R}_H$  is topologically equivalent to  $\mathbb{R}_+^{n+1}$ , so theorem 1 applies to these subsets individually. This means that we can represent preferences within each  $\mathbb{R}_H$  by reference-independent utility functions

$$U_H(\mathbf{x}) = \text{sign}(\rho_H) \left[ \delta_{0,H} x_0^{\rho_H} + \sum_{i \in H} \delta_{a,H} x_i^{\rho_H} + \sum_{i \in N/H} \delta_{b,H} x_i^{\rho_H} \right]$$

such that  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}_H : [\mathbf{x} \succeq \mathbf{y} \longleftrightarrow U_H(\mathbf{x}) \geq U_H(\mathbf{y})]$ .<sup>23</sup> Neutrality implies that a permutation of other individual's payoffs does not affect preferences, which rules out differences in the weights  $\delta$  across individuals  $i$ , although the weights may be different depending on the position of  $x_i$

<sup>22</sup>The rationale for this particular normalisation is to allow for negative weights. If the weights on individual payoffs are negative, the weighted sum inside the brackets could be negative. The chosen normalisation avoids the problem of raising a negative number to a non-integer power, and thus represents the preference relation in question with more generality.

<sup>23</sup>Or the equivalent Cobb-Douglas form. We omit this specification in what follows, which should be understood as always including the Cobb-Douglas form as the limiting case as the exponent in the utility function goes to zero.

relative to  $x_0$ . Non-discrimination, furthermore, implies that the *identity* the individuals in  $H$  is irrelevant for preferences. This means that the same utility function must represent preferences across all the subsets  $\mathbb{R}_H$  for which the cardinality of  $H$  is the same. The utility function is therefore unique (up to a monotonic transformation) within each subset of distributions  $\mathbb{R}_h \equiv \{\mathbf{x} \in \mathbb{R}_+^{n+1} : h = |\{i \in N : x_i \leq x_0\}|\}$ , defined by the number of individuals ranked above or below the decision-maker. This permits us to represent preferences with the utility function:

$$U_h(\mathbf{x}) = \text{sign}(\rho_h) \left[ \delta_{0,h} x_0^{\rho_h} + \sum_{x_i \neq 0 \leq x_0} \delta_{a,h} x_i^{\rho_h} + \sum_{x_i > x_0} \delta_{b,h} x_i^{\rho_h} \right],$$

with  $h = |\{i \in N : x_i \leq x_0\}|$

This is a separate CES-RI utility function for each partition of the payoff set defined by the number of individuals with a payoff lower or higher than that of the decision-maker.

By Debreu (1954), we know that the closedness of the at-least-as-good-as set and the no-better-than set entail that the utility function  $U(\mathbf{x})$  is continuous on its domain  $\mathbb{R}_+^{n+1}$ . That is, if  $\{\mathbf{y}_m\}_{m=1}^\infty$  is a sequence in  $\mathbb{R}_+^{n+1}$  such that  $\lim_{m \rightarrow \infty} \mathbf{y}_m = \mathbf{x} \in \mathbb{R}_+^{n+1}$ , then  $\lim_{m \rightarrow \infty} U(\mathbf{y}_m) = U(\mathbf{x})$ .

Consider an allocation  $\mathbf{x} \in \mathbb{R}_{h+1}$  for some  $0 \leq h < n$ , such that  $x_j = x_0$  and  $x_{i \neq j} \neq x_0$  for some  $j \in N$ . The relevant function representing preferences at this point is  $U_{h+1}(\mathbf{x})$ . Now take a point such that  $\mathbf{y}_{\sim j} = \mathbf{x}_{\sim j}$  and  $y_j > x_j$ ; clearly  $\mathbf{y} \in \mathbb{R}_h$ , and so the relevant utility function at this point is  $U_h(\mathbf{y})$ . Consider the sequence  $\{\mathbf{y}_m\}_{m=1}^\infty$  defined by  $\mathbf{y}_m = \{x_0, x_1, \dots, x_j + (y_j - x_j)/m, \dots, x_n\}$  for all  $m > 0$ . Clearly this sequence converges to  $\mathbf{x} \in \mathbb{R}_{h+1}$ . Nevertheless, since  $\mathbf{y}_m \in \mathbb{R}_h$  for all  $m > 0$ , we have  $U(\mathbf{y}_m) = U_h(\mathbf{y}_m)$  as the relevant utility function at  $\mathbf{y}_m$  for all finite  $m > 0$ . The continuity of  $U$  then implies:

$$(14) \quad \lim_{m \rightarrow \infty} U_h(\mathbf{y}_m) = U_{h+1}(\mathbf{x})$$

or

$$\begin{aligned}
 (15) \quad & \text{sign}(\rho_h) \left[ \delta_{0,h} x_0^{\rho_h} + \delta_{b,h} \left[ x_j + \lim_{m \rightarrow \infty} \left( \frac{y_j - x_j}{m} \right) \right]^{\rho_h} + \sum_{x_i \neq 0 \leq x_0} \delta_{a,h} x_i^{\rho_h} + \sum_{x_i \neq j > x_0} \delta_{b,h} x_i^{\rho_h} \right] \\
 & = \text{sign}(\rho_{h+1}) \left[ (\delta_{0,h+1} + \delta_{a,h+1}) x_0^{\rho_{h+1}} + \sum_{x_i \notin \{0,j\} \leq x_0} \delta_{a,h+1} x_i^{\rho_{h+1}} + \sum_{x_i \neq j > x_0} \delta_{b,h+1} x_i^{\rho_{h+1}} \right].
 \end{aligned}$$

Substituting for the limit on the left-hand side and simplifying yields

$$\begin{aligned}
 (16) \quad & \text{sign}(\rho_h) \left[ (\delta_{0,h} + \delta_{b,h}) x_0^{\rho_h} + \sum_{x_i \notin \{0,j\} \leq x_0} \delta_{a,h} x_i^{\rho_h} + \sum_{x_i \neq j > x_0} \delta_{b,h} x_i^{\rho_h} \right] \\
 & = \text{sign}(\rho_{h+1}) \left[ (\delta_{0,h+1} + \delta_{a,h+1}) x_0^{\rho_{h+1}} + \sum_{x_i \notin \{0,j\} \leq x_0} \delta_{a,h+1} x_i^{\rho_{h+1}} + \sum_{x_i \neq j > x_0} \delta_{b,h+1} x_i^{\rho_{h+1}} \right]
 \end{aligned}$$

which must be true for any  $\mathbf{x} \in \mathbb{R}_{h+1} \cap Cl(\mathbb{R}_h)$ . This equality can only hold generally if the exponents are the same on both sides.<sup>24</sup> We may therefore write  $\rho_h = \rho, \forall h \in N$ . Collecting terms simplifies the condition to:

$$\begin{aligned}
 (17) \quad & [(\delta_{0,h} + \delta_{b,h}) - (\delta_{0,h+1} + \delta_{a,h+1})] x_0^\rho \\
 & = (\delta_{a,h+1} - \delta_{a,h}) \sum_{x_i \notin \{0,j\} \leq x_0} x_i^\rho + (\delta_{b,h+1} - \delta_{b,h}) \sum_{x_i \neq j > x_0} x_i^\rho
 \end{aligned}$$

which, because of the arbitrary choice of  $\mathbf{x}$ , entails:

$$(18) \quad \begin{bmatrix} (\delta_{a,h+1} - \delta_{a,h}) \\ (\delta_{b,h+1} - \delta_{b,h}) \\ (\delta_{0,h} + \delta_{b,h}) - (\delta_{0,h+1} + \delta_{a,h+1}) \end{bmatrix} = 0$$

<sup>24</sup>In the trivial case where  $\delta_{0,h} = \delta_{a,h} = \delta_{b,h} = 0$ , the exponents do not affect the constant null value of the function. In that case also we may therefore impose  $\rho_h = \rho, \forall h \in N_0$ .

and it follows that, for all  $h \in N_0$ :

$$(19) \quad \begin{aligned} \delta_{a,h} &= \delta_a \\ \delta_{b,h} &= \delta_b \\ \delta_{0,h} &= \delta_{0,n} + (n-h)(\delta_a - \delta_b) \end{aligned}$$

We have proved that preferences can be represented by the utility function:

$$U(\mathbf{x}) = \text{sign}(\rho) \left[ [\delta_{0,n} + (n-h)(\delta_a - \delta_b)] x_0^\rho + \delta_a \sum_{x_i \neq 0 \leq x_0} x_i^\rho + \delta_b \sum_{x_i > x_0} x_i^\rho \right],$$

with  $h \equiv |i \in N_0 : x_i \leq x_0|$ .

To arrive at the form in the theorem, normalise the utility function by the transformation  $T(z) \equiv z / [\delta_{0,n} + n\delta_a]$ , which is monotonic, since by axiom MI we have  $\delta_{0,n} + n\delta_a > 0$ . Define  $\alpha \equiv T(\delta_a)$  and  $\beta \equiv T(\delta_b)$  and note that  $T(\delta_{0,n}) = 1 - n\alpha$  to write the utility function as

$$U(\mathbf{x}) = \text{sign}(\rho) \left[ (1 - h\alpha - (n-h)\beta) x_0^\rho + \alpha \sum_{x_i \neq 0 \leq x_0} x_i^\rho + \beta \sum_{x_i > x_0} x_i^\rho \right]$$

with  $h \equiv |i \in N_0 : x_i \leq x_0|$

which completes the proof. □

**A.3. Proofs of Theorems 3, 4 and 5.** It is clear by inspection that the functional form in theorem 3 satisfies the axioms. We here prove the necessity of that functional form. By a similar reasoning to the proof of theorem 1, axioms S and H imply that for any given  $\mathbf{r}$ , the utility function  $U_{\mathbf{r}}(\bullet)$  is of the CES form. Thus the preference structure can be represented by any monotonic transformation of:

$$U_{\mathbf{r}}(\mathbf{x}) = \begin{cases} \text{sign}(\rho(\mathbf{r})) \left[ \delta_0(\mathbf{r}) x_0^{\rho(\mathbf{r})} + \sum_{i=1}^n \delta_i(\mathbf{r}) x_i^{\rho(\mathbf{r})} \right] & \text{if } \rho \neq 0 \\ \text{or} \\ \delta_0(\mathbf{r}) \ln x_0 + \sum_{i=1}^n \delta_i(\mathbf{r}) \ln x_i \end{cases}$$

RN entails that the functions  $\delta_i(\bullet)$  must be identical for all  $i \in N$ , so we can write  $\delta_i(r_i, \mathbf{r}_{\sim i}) = \delta_a(r_i, \mathbf{r}_{\sim i}), \forall i \in N$ .

We establish two lemmata before proving the theorem:

**Lemma 2.**  $\rho$  is independent of  $\mathbf{r}$ .

*Proof.* Choose distinct  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  such that for some  $\{i, j\} \in N$  we have  $\mathbf{x}_{\sim\{i,j\}} = \mathbf{y}_{\sim\{i,j\}}$ . Choose a reference point  $\mathbf{r} \in \mathbb{R}_+^{n+1}$  such that  $\mathbf{x} \sim_{\mathbf{r}} \mathbf{y}$ . Then it must be the case that

$$(20) \quad \delta_a(r_i, \mathbf{r}_{\sim i}) x_i^{\rho(\mathbf{r})} + \delta_a(r_j, \mathbf{r}_{\sim j}) x_j^{\rho(\mathbf{r})} = \delta_a(r_i, \mathbf{r}_{\sim i}) y_i^{\rho(\mathbf{r})} + \delta_a(r_j, \mathbf{r}_{\sim j}) y_j^{\rho(\mathbf{r})}.$$

Assuming  $\delta_a(r_j, \mathbf{r}_{\sim j}) \neq 0$ , we may rewrite this equality as

$$(21) \quad \frac{\delta_a(r_i, \mathbf{r}_{\sim i})}{\delta_a(r_j, \mathbf{r}_{\sim j})} = -\frac{x_j^{\rho(\mathbf{r})} - y_j^{\rho(\mathbf{r})}}{x_i^{\rho(\mathbf{r})} - y_i^{\rho(\mathbf{r})}}.$$

Consider a change in the reference payoff of an individual  $k \in N \setminus \{i, j\}$ . Axiom WIIR (and, *a fortiori*, axiom IIR) requires that the indifference between  $\mathbf{x}$  and  $\mathbf{y}$  should be unaffected. Therefore,

$$(22) \quad \frac{d}{dr_k} \left( \frac{\delta_a(r_i, \mathbf{r}_{\sim i})}{\delta_a(r_j, \mathbf{r}_{\sim j})} \right) = -\frac{\partial}{\partial \rho} \left( \frac{x_j^{\rho(\mathbf{r})} - y_j^{\rho(\mathbf{r})}}{x_i^{\rho(\mathbf{r})} - y_i^{\rho(\mathbf{r})}} \right) \frac{d\rho}{dr_k}$$

which can only be generally true if  $d\rho/dr_k = 0$ , since the right-hand side otherwise varies with  $\mathbf{x}_{\{i,j\}}$  and  $\mathbf{y}_{\{i,j\}}$ , while the left-hand side does not. The same argument can be made for any  $k \in N$ .  $\rho$  can therefore at most depend on  $r_0$ .

Now consider a change in  $r_0$ , the fair claim of the decision-maker herself. To see that  $d\rho/dr_0$  must be zero, choose another allocation  $\mathbf{z}$ , distinct from  $\mathbf{x}$  and  $\mathbf{y}$  but satisfying  $\mathbf{z}_{\sim\{i,j\}} = \mathbf{x}_{\sim\{i,j\}} = \mathbf{y}_{\sim\{i,j\}}$  and  $\mathbf{x} \sim_{\mathbf{r}} \mathbf{y} \sim_{\mathbf{r}} \mathbf{z}$ . The following relationship then holds:

$$(23) \quad \frac{\delta_a(r_i, \mathbf{r}_{\sim i})}{\delta_a(r_j, \mathbf{r}_{\sim j})} = \frac{x_j^{\rho(r_0)} - y_j^{\rho(r_0)}}{x_i^{\rho(r_0)} - y_i^{\rho(r_0)}} = \frac{x_j^{\rho(r_0)} - z_j^{\rho(r_0)}}{x_i^{\rho(r_0)} - z_i^{\rho(r_0)}} = \frac{y_j^{\rho(r_0)} - z_j^{\rho(r_0)}}{y_i^{\rho(r_0)} - z_i^{\rho(r_0)}}.$$

By axiom RH, preference rankings over allocations are unaffected by identical equiproportional scalings of the fairness claims, so the same relationship must hold for any positive scaling factor  $\lambda > 0$ :

$$(24) \quad \frac{x_j^{\rho(\lambda r_0)} - y_j^{\rho(\lambda r_0)}}{x_i^{\rho(\lambda r_0)} - y_i^{\rho(\lambda r_0)}} = \frac{x_j^{\rho(\lambda r_0)} - z_j^{\rho(\lambda r_0)}}{x_i^{\rho(\lambda r_0)} - z_i^{\rho(\lambda r_0)}} = \frac{y_j^{\rho(\lambda r_0)} - z_j^{\rho(\lambda r_0)}}{y_i^{\rho(\lambda r_0)} - z_i^{\rho(\lambda r_0)}}, \forall \lambda > 0.$$

But this can only be generally true if  $d\rho/dr_0 = 0$ . It follows that  $\rho$  is constant.  $\square$

**Lemma 3.**  $\delta_a(r_i, \mathbf{r}_{\sim i})$  and  $\delta_0(\mathbf{r})$  are homogeneous of the same degree.

*Proof.* Consider two allocations  $\mathbf{x}, \mathbf{y} \in \mathbb{R}_{++}^{n+1}$  satisfying  $\mathbf{x} \sim_{\mathbf{r}} \mathbf{y}$ , so that  $U_{\mathbf{r}}(\mathbf{x}) = U_{\mathbf{r}}(\mathbf{y})$ . Axiom RH implies that with  $\mathbf{s} = \lambda \mathbf{r}, \forall \lambda > 0$ , we have:

$$(25) \quad U_{\mathbf{s}}(\mathbf{x}) = U_{\mathbf{s}}(\mathbf{y})$$

and by applying lemma 2, we may write:

$$(26) \quad \delta_0(\lambda \mathbf{r}) x_0^\rho + \sum_{i=1}^n \delta_a(\lambda r_i, \lambda \mathbf{r}_{\sim i}) x_i^\rho = \delta_0(\lambda \mathbf{r}) y_0^\rho + \sum_{i=1}^n \delta_a(\lambda r_i, \lambda \mathbf{r}_{\sim i}) y_i^\rho.$$

Taking the derivative of both sides of equation 26 with respect to  $\lambda$  and evaluating at  $\lambda = 1$ , we get

$$(27) \quad \begin{aligned} & \left( \sum_{k=0}^n \frac{\partial \delta_0(\mathbf{r})}{\partial r_k} r_k \right) x_0^\rho + \sum_{i=1}^n \left( \sum_{k=0}^n \frac{\partial \delta_a(r_i, \mathbf{r}_{\sim i})}{\partial r_k} r_k \right) x_i^\rho \\ &= \left( \sum_{k=0}^n \frac{\partial \delta_0(\mathbf{r})}{\partial r_k} r_k \right) y_0^\rho + \sum_{i=1}^n \left( \sum_{k=0}^n \frac{\partial \delta_a(r_i, \mathbf{r}_{\sim i})}{\partial r_k} r_k \right) y_i^\rho \end{aligned}$$

This equality holds for any  $\mathbf{x}, \mathbf{y}, \mathbf{r}$  that satisfy  $\mathbf{x} \sim_{\mathbf{r}} \mathbf{y}$ . Now move the point  $\mathbf{x}$  along the utility contour defined by  $\mathbf{r}$  in  $(x_i, x_j)$ -space ( $i, j \in N$ ). Since the point  $\mathbf{y}$  is not moving, we have:

$$(28) \quad \rho \left( \sum_{k=0}^n \frac{\partial \delta_a(r_i, \mathbf{r}_{\sim i})}{\partial r_k} r_k \right) x_i^{\rho-1} + \rho \left( \sum_{k=0}^n \frac{\partial \delta_a(r_j, \mathbf{r}_{\sim j})}{\partial r_k} r_k \right) x_j^{\rho-1} \frac{dx_j}{dx_i} \Big|_{dU_{\mathbf{r}}(\mathbf{x})=0, dx_{-\{i,j\}}=0} = 0$$

which simplifies to

$$(29) \quad \frac{\left( \sum_{k=0}^n \frac{\partial \delta_a(r_i, \mathbf{r}_{\sim i})}{\partial r_k} r_k \right)}{\delta_a(r_i, \mathbf{r}_{\sim i})} = \frac{\left( \sum_{k=0}^n \frac{\partial \delta_a(r_j, \mathbf{r}_{\sim j})}{\partial r_k} r_k \right)}{\delta_a(r_j, \mathbf{r}_{\sim j})} = c$$

(with  $\delta_a(r_i, \mathbf{r}_{\sim i})$  and  $\delta_a(r_j, \mathbf{r}_{\sim j})$  nonzero), where the equality to a constant  $c$  follows because the first equality holds for any  $\{i, j\} \in N$ . We have established that  $\delta_a$  is homogeneous of degree  $c$ . The same conclusion can be established for  $\delta_0$  by moving the point  $\mathbf{x}$  in  $(x_0, x_i)$ -space with an arbitrary choice of  $i \in N$ .  $\square$

*Proof of theorem 3.* Sufficiency of the functional form is clear from inspection. For necessity, note that IIR implies WIIR, so we prove the weaker case first. As before, choose  $\{i, j\} \in N$ , and

$\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^{n+1}$  such that  $\mathbf{x}_{\sim\{i,j\}} = \mathbf{y}_{\sim\{i,j\}}$ . Choose a reference point  $\mathbf{r}$  such that  $\mathbf{x} \sim_{\mathbf{r}} \mathbf{y}$ , which implies that:

$$(30) \quad \delta_a(r_i, \mathbf{r}_{\sim i})(x_i^\rho - y_i^\rho) = -\delta_a(r_j, \mathbf{r}_{\sim j})(x_j^\rho - y_j^\rho).$$

Now by WIIR this equality has to hold independently of the value of  $\mathbf{r}_{\sim\{i,j\}}$ , and for each  $i \in N$  it must hold for any arbitrarily chosen  $j \in N \setminus \{i\}$ . Therefore, it must be possible to factor out each individual's reference point separately from each side of this equation, except the reference point of the decision-maker. It follows that the  $\delta_a$ -function must have the form

$$(31) \quad \delta_a(r_i, \mathbf{r}_{\sim i}) = \phi(r_i, r_0) \prod_{j=1}^n \psi(r_j)$$

(whenever  $\delta_a \neq 0$ ). A similar argument shows that

$$(32) \quad \delta_0(r_0, \mathbf{r}_{\sim 0}) = \phi_0(r_0) \prod_{j=1}^n \psi(r_j)$$

(whenever  $\delta_0 \neq 0$ ). Given that the  $\delta$ -functions are homogenous of the same degree  $c$ ,  $\phi_0(r_0)$  and  $\phi(r_i, r_0)$  must both be homogenous of the same degree  $\gamma = c - n\theta$ , where  $\theta$  is the degree of homogeneity of the functions  $\psi(r_j)$ . We may therefore write  $\phi_0(r_0) = r_0^\gamma$  (multiplying any linear coefficient into the other factors on the right-hand side of equation 32) and  $\phi(r_i, r_0) = r_0^\gamma \phi\left(\frac{r_i}{r_0}, 1\right)$ .

Define the univariate function  $\phi(r_i/r_0) \equiv \phi\left(\frac{r_i}{r_0}, 1\right)$ , and denote by  $\Phi(\mathbf{r}) \equiv \sum_{i=1}^n \phi(r_i/r_0)$ . Axiom MI entails that  $r_0^\gamma [\prod_{i=1}^n \psi(r_i)] [1 + \Phi(\mathbf{r})] \geq 0$ , so  $\text{sign} [\prod_{i=1}^n \psi(r_i)] = \text{sign} [1 + \Phi(\mathbf{r})]$ . Given the arbitrary choice of  $\mathbf{r}$ , this is only possible if either  $\Phi(\mathbf{r}) < -1$  or  $\Phi(\mathbf{r}) > -1$ , which implies

$$\text{either } \max_{z \geq 0} \phi(z) < -1/n \text{ or } \min_{z \geq 0} \phi(z) > -1/n$$

as stated in the theorem.

We may normalise the utility function by  $r_0^\gamma |\prod_{i=1}^n \psi(r_i)|$  to represent preferences by  $\tilde{U}_{\mathbf{r}}(\mathbf{x}) = \text{sign} [\rho(1 + \Phi(\mathbf{r}))] [x_0^\rho + \sum_{i=1}^n \phi(r_i/r_0) x_i^\rho]$ . Now consider any  $\mathbf{x}, \mathbf{y}, \mathbf{r} \in \mathbb{R}_+^{n+1}$  such that  $\mathbf{x} \sim_{\mathbf{r}} \mathbf{y}$ . Then  $\tilde{U}_{\mathbf{r}}(\mathbf{x}) = \tilde{U}_{\mathbf{r}}(\mathbf{y})$ , and given the continuity of the preference structure, axiom RPI entails that

for any  $j \in N_0$ :

$$(33) \quad \frac{d}{dr_j} [\tilde{U}_{\mathbf{r}}(\mathbf{x}) - \tilde{U}_{\mathbf{r}}(\mathbf{y})] \geq 0$$

if and only if  $x_j - y_j \geq 0$ .

In this case we have:

$$(34) \quad \frac{d}{dr_j} [\tilde{U}_{\mathbf{r}}(\mathbf{x}) - \tilde{U}_{\mathbf{r}}(\mathbf{y})] = \text{sign}[\rho(1 + \Phi(\mathbf{r}))] \sum_{i=1}^n \frac{d\phi(r_i/r_0)}{dr_j} (x_i^\rho - y_i^\rho)$$

$$(35) \quad = \text{sign}[\rho(1 + \Phi(\mathbf{r}))] \frac{\phi'(r_j/r_0)}{r_0} (x_j^\rho - y_j^\rho), \quad \forall j \in N$$

The expression in the previous line satisfies the sign condition in equation 33 if and only if  $\phi'(r_i/r_0)$  has the same sign as  $(1 + \Phi(\mathbf{r}))$ .

Finally, we divide  $\tilde{U}_{\mathbf{r}}(\cdot)$  by  $|1 + \Phi(\mathbf{r})|$  to arrive at the functional form in theorem 3:

$$\begin{aligned} U_{\mathbf{r}}(\mathbf{x}) &= \tilde{U}_{\mathbf{r}}(\mathbf{x}) / |1 + \Phi(\mathbf{r})| \\ &= \text{sign}(\rho) \left[ \left( 1 - \sum_{i=1}^n \alpha_i(\mathbf{r}) \right) x_0^\rho + \sum_{i=1}^n \alpha_i(\mathbf{r}) x_i^\rho \right] \\ \text{where } \alpha_i(\mathbf{r}) &= \frac{\phi(r_i/r_0)}{1 + \Phi(\mathbf{r})}. \end{aligned}$$

This completes the proof. □

*Proof of theorem 4.* Since IIR entails WIIR, all the reasoning in the proof of theorem 3 still holds. Replacing WIIR by IIR additionally allows us to factor out the decision-maker's reference point  $r_0$  from both sides of equation 30. If the reference payoff of any individual including the decision-maker can be factored out, it must be possible to write the function  $\phi(r_i, r_0)$  in equation 31 as a product of univariate functions:

$$(36) \quad \phi(r_i, r_0) = \xi(r_i) \psi_0(r_0).$$

The homogeneity of degree  $c - n\theta$  of the function  $\phi(r_i, r_0)$  implies that  $\xi(r_i)$  and  $\psi_0(r_0)$  are monomials whose exponents sum to  $c - n\theta$ . Define  $\gamma$  and  $\zeta$  as the degrees of the monomials  $\xi(r_i)$  and  $\psi_0(r_0)$ , respectively; we have  $\gamma + \zeta = c - n\theta$ . We can therefore write  $\phi_0(r_0) = r_0^{\gamma+\zeta}$  and  $\phi(r_i, r_0) = ar_0^{\gamma+\zeta}(r_i/r_0)^\gamma$ , where  $a$  is a constant. The univariate function  $\phi(r_i/r_0)$  can now be specified; it is  $\phi(r_i/r_0) = a(r_i/r_0)^\gamma$ . Substituting this expression into the functional form derived in theorem 3<sup>25</sup> yields the desired result. The boundedness of  $\phi(r_i/r_0)$  derived in the proof of theorem 3 entails that  $a \geq 0, \gamma \geq 0$ .  $\square$

*Proof of theorem 5.* Theorem 5 follows from a straightforward combination of the proof of theorem 2 and the proofs in this subsection.  $\square$

#### REFERENCES

- Andreoni, James, & Miller, John. 2002. Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica*, **70**(2), 737–753.
- Andreoni, James, & Petrie, Ragan. 2004. *Beauty, Gender and Stereotypes: Evidence From Laboratory Experiments*. Unpublished manuscript.
- Andreoni, James, Brown, Paul M., & Vesterlund, Lise. 2002. What Makes an Allocation Fair? Some Experimental Evidence. *Games and Economic Behavior*, **40**, 1–24.
- Arrow, Kenneth J., Chenery, H. B., Minhas, B. S., & Solow, R. M. 1961. Capital-Labor Substitution and Economic Efficiency. *The Review of Economics and Statistics*, **XLIII**(3), 225–250.
- Bateman, Ian, Munro, Alistair, Rhodes, Bruce, Starmer, Chris, & Sugden, Robert. 1997. A Test of the Theory of Reference-Dependent Preferences. *The Quarterly Journal of Economics*, **112**, 479–505.
- Blackorby, Charles, & Donaldson, David. 1982. Ratio-scale and translation-scale full interpersonal comparability without domain restrictions: Admissible social-evaluation functions. *International Economic Review*, **23**(2), 249–267.
- Bolton, Gary E, & Ockenfels, Axel. 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review*, **90**(1), 166–193.

<sup>25</sup>The monotonic transformation used to normalise the utility function is now written  $T(z) = z/r_0^{\gamma+\zeta} \left| \prod_{j=1}^n \psi(r_j) \right|$ , since the degree of homogeneity of the function  $\phi(r_i, r_0)$  is now denoted by  $\gamma + \zeta$ .

- Brandts, Jordi, & Charness, Gary. 1999. *Retribution in a Cheap-Talk Experiment*. Unpublished manuscript.
- Brandts, Jordi, & Sola, Carles. 2001. Reference Points and Negative Reciprocity in Simple Sequential Games. *Games and Economic Behavior*, **36**(2), 138.
- Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. The Roundtable Series in Behavioral Economics. New York and Princeton: Princeton University Press and Russell Sage Foundation.
- Charness, Gary, & Dufwenberg, Martin. 2006. Promises and Partnership. *Econometrica*, **74**(6), 1579–1601.
- Charness, Gary, & Rabin, Matthew. 2002. Understanding Social Preferences With Simple Tests. *The Quarterly Journal of Economics*, **117**, 817–869.
- Chew, Soo Hong, & Epstein, Larry G. 1989. A Unifying Approach to Axiomatic Non-expected Utility Theories. *Journal of Economic Theory*, **49**, 207–240.
- Cox, James C., Friedman, Daniel, & Gjerstad, Steven. 2007. A Tractable Model of Reciprocity and Fairness. *Games and Economic Behavior*, **59**(1), 17–45.
- Debreu, Gérard. 1954. Representation of a Preference Ordering by a Numerical Function. *Pages 159–165 of: Thrall, Robert MacDowell, Davis, C. H., & Coombs, R. L. (eds), Decision Processes*. New York: John Wiley.
- Debreu, Gérard. 1983. Topological Methods in Cardinal Utility Theory. *Pages 16–26 of: Arrow, Kenneth J., Karlin, Samuel, & Suppes, Patrick (eds), Mathematical Methods in the Social Sciences*. Stanford mathematical studies in the social sciences, vol. 4. Stanford, CA: Stanford University Press.
- Dryzek, John S., & List, Christian. 2003. Social Choice Theory and Deliberative Democracy: A Reconciliation. *British Journal of Political Science*, **33**(1), 1–28.
- Dufwenberg, Martin, & Kirchsteiger, Georg. 2004. A Theory of Sequential Reciprocity. *Games and Economic Behavior*, **47**(2), 268–298.
- Falk, Armin, & Fischbacher, Urs. 2000. A Theory of Reciprocity. *Games and Economic Behavior*, **54**(2), 293–315.

- Fehr, Ernst, & Schmidt, Klaus M. 1999. A Theory of Fairness, Competition and Cooperation. *The Quarterly Journal of Economics*, **114**, 817–868.
- Fehr, Ernst, & Schmidt, Klaus M. 2001 (February). *Theories of Fairness and Reciprocity - Evidence and Economic Applications*. CEPR Discussion Paper 2703.
- Forsythe, Robert, Horowitz, Joel L., Savin, N. E., & Sefton, Martin. 1994. Fairness in Simple Bargaining Experiments. *Games and Economic Behavior*, **6**, 347–369.
- Frey, Bruno S., Benz, Matthias, & Stützer, Alois. 2004. Introducing Procedural Utility: Not only What, but also How matters. *Journal of Institutional and Theoretical Economics*, **160**(3), 377–401.
- Gächter, Simon, & Riedl, Arno. 2005. Moral Property Rights in Bargaining with Infeasible Claims. *Management Science*, **51**(2), 249–263.
- Güth, Werner, Schmittberger, R., & Schwarze, B. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, **3**(4), 367–388.
- Güth, Werner, Huck, Steffen, & Müller, Wieland. 2001. The Relevance of Equal Splits in Ultimatum Games. *Games and Economic Behavior*, **37**, 161–169.
- Kalai, E., & Smorodinsky, M. 1975. Other solutions to Nash's bargaining problem. *Econometrica*, **43**, 513–18.
- Karni, Edi, & Safra, Zvi. 2002. Individual Sense of Justice: A Utility Representation. *Econometrica*, **70**(1), 263–284.
- Ledyard, John. 1995. Public Goods: A Survey of Experimental Research. *Pages 111–194 of: Kagel, John H., & Roth, Alvin E. (eds), The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
- McFadden, Daniel. 1963. Constant Elasticity of Substitution Production Functions. *The Review of Economic Studies*, **30**(2), 73–83.
- Munro, Alistair, & Sugden, Robert. 2003. On the theory of reference-dependent preferences. *Journal of Economic Behavior and Organization*, **50**(4), 407–428.
- Neilson, William S. 2002 (January). *An Axiomatic Characterization of the Fehr-Schmidt Model of Inequity Aversion*. Unpublished manuscript.

- Neilson, William S. 2006. Axiomatic reference-dependence in behavior toward others and toward risk. *Economic Theory*, **28**(3), 681–692.
- Nozick, Robert. 1974. *Anarchy, State and Utopia*. New York: Basic Books.
- Ok, Efe A., & Koçkesen, Levent. 2000. Negatively Interdependent Preferences. *Social Choice and Welfare*, **17**(3), 533–558.
- Pattanaik, Prasanta K., & Suzumura, Kotaro. 1994. Rights, Welfarism and Social Choice. *The American Economic Review*, **84**(2), 435–439.
- Plott, Charles. 1973. Path Independence, Rationality, and Social Choice. *Econometrica*, **41**(6), 1075–1091.
- Prasnikar, Vesna, & Roth, Alvin E. 1992. Considerations of Fairness and Strategy: Experimental Data from Sequential Games. *The Quarterly Journal of Economics*, **107**, 865–888.
- Rabin, Matthew. 1993. Incorporating Fairness into Game Theory and Economics. *The American Economic Review*, **83**(5), 1281–1302.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Roth, Alvin E., Prasnikar, Vesna, Okuno-Fujiwara, Masahiro, & Zamir, Shmuel. 1991. Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *The American Economic Review*, **81**(5), 1068–1095.
- Sandbu, Martin Eiliv. 2003. *Explorations in Process-Dependent Preference Theory*. Ph.D. thesis, Harvard University.
- Sandbu, Martin Eiliv. 2007. Fairness and the roads not taken: An experimental test of non-reciprocal set-dependence in distributive preferences. *Games and Economic Behavior*, **61**(1), 113–130.
- Sen, Amartya. 1995. Rationality and Social Choice. *The American Economic Review*, **85**(1), 1–24.
- Sen, Amartya. 1997. Maximization and the Act of Choice. *Econometrica*, **65**(4), 745–779.
- Sen, Amartya. 1999. *Development as Freedom*. New York: Anchor Books.
- Suzumura, Kotaro. 1999. Consequences, opportunities and procedures. *Social Choice and Welfare*, **16**(1), 17–40.

- Tversky, Amos, & Kahneman, Daniel. 1991. Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics*, **106**(4), 1039–1061.
- Uzawa, Hirofumi. 1962. Production Functions with Constant Elasticities of Substitution. *The Review of Economic Studies*, **29**(4), 291–299.

DEPARTMENT OF LEGAL STUDIES AND BUSINESS ETHICS, WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA, 3730 WALNUT STREET, PHILADELPHIA, PA 19104

*E-mail address:* sandbu@post.harvard.edu