Interdependent Altruistic Preference Models

Jay Simon
American University
jaysimon@american.edu

Donald Saari
University of California, Irvine

L. Robin Keller
University of California, Irvine
LRKeller@uci.edu

Altruistic preferences, or the desire to improve the well-being of others even at one's own expense, can be difficult to incorporate into traditional value and utility models. It is straightforward to construct a multi-attribute preference structure for one decision maker that includes the outcomes experienced by others. However, when multiple individuals incorporate one another's well-being into their decision making, this creates complex interdependencies that must be resolved before the preference models can be applied. We provide representation theorems for additive altruistic value functions for two-person, n-person, and group outcomes in which multiple individuals are altruistic. We find that in most cases it is possible to resolve the preference interdependencies and that modeling the preferences of altruistic individuals and groups is tractable.

Key words: altruism, preferences, decision theory

1. Introduction

The economic concept of altruism refers to a person making a decision to increase the well-being of others despite incurring some individual cost. The existence of altruism challenges the traditional definition of utility over wealth or bundles of goods experienced as an individual, since an altruistic person's choices can violate axioms of single-attribute utility maximization (Savage 1954). Altruism, however, is not incompatible with models of rational choice; it simply requires a multi-attribute function that can incorporate the preferences of other individuals, and is well suited to decision analysis techniques.

In this paper, we focus on ordinal value functions, though the approach could be adapted to utility functions over gambles with the development of suitable representation theorems. To represent altruistic preferences, we will distinguish between a *traditional* single-attribute value function over a decision maker's individual outcomes that disregards the outcomes of others (what Bell and Keeney (2009) call an *egotistical* function), and a multi-attribute value function that is permitted to contain another individual's value as an attribute. In particular, we allow an individual's altruistic value function to contain another individual's altruistic value rather than traditional value. That is, if two altruistic individuals go out to dinner together, the altruistic component of each one's value function is permitted to include not only the other individual's enjoyment of the meal, but also anything else that contributes to the other individual's value function, including altruism.

However, modeling altruistic preferences in this way creates both mathematical and philosophical challenges. Two individuals who are altruistic toward one another will have an interdependent pair of value functions. If Person 1's value depends on Person 2's value, which depends on Person 1's value, which depends on Person 2's value, etc., how can this potentially infinite sequence be resolved? To overcome this hurdle and incorporate altruism in this case, it must be possible for the individuals' altruistic values to be consistent with one another, and for the sake of prescriptive application, closed-form value functions are desirable. We will accomplish this in two steps. First, we will develop representation theorems for altruistic value functions that contain the other individual's traditional values as parameters. We examine additive value functions as an example; other analogous representation theorems could be developed for other forms. Second, we will provide conditions under which a set of altruistic value functions containing the other individual's altruistic values as parameters can be expressed in the simpler form to which the representation theorems apply.

There are two advantages to this approach. The inclusion of another individual's either traditional value or altruistic value in a value function would be considered *non-paternalistic* altruism (as opposed to *paternalistic* altruism, which incorporates others' outcomes but not their preferences). We believe that using altruistic value as a parameter is truer to the spirit of non-paternalistic altruism, in that it captures the other individual's preferences as accurately as possible. In addition, it may be more natural to assess tradeoffs regarding another individual's overall well-being rather than another individual's value achieved from a single attribute; we will discuss this in Section 4.

We will then expand the approach to consider more than two individuals. While the theory generalizes naturally, some of the conditions become more difficult to check in practice, and assessment is more challenging. However, should the individuals be altruistic due to their belonging to a common group, an additional preference condition can greatly simplify both modeling and preference elicitation. We will also demonstrate the application of these altruistic multi-attribute value functions using simple illustrative examples.

The remainder of the paper proceeds as follows. Section 2 provides additional background and discusses related prior work. In Section 3.1, we consider the case of two people, Section 3.2 considers n (more than 2) people, and Section 3.3 considers the addition of a "group outcome." Section 4 discusses the assessment of altruistic value functions. Section 5 presents two examples of decisions involving altruistic preferences. Finally, Section 6 concludes the paper.

2. Background

Even in the simple case that considers only one individual's preferences, it is not immediately obvious that an altruistic value function exists for that individual. The conditions typically used to justify the existence of a value function are rather strong when altruism is included. Simon (2016) provides representation theorems justifying the use of altruistic preference functions in the two-person case that rely on weaker conditions; the current paper extends these results to settings in which interdependent altruism exists.

In contrast to our models, Bell and Keeney (2009) limit their novel altruistic preference model to include only the traditional value of the outcomes to multiple individuals. In their Result 4, they argue that only traditional values should be used, thus avoiding a problem of double counting and infinite comparisons. We resolve that problem by using the implicit function theorem. We will discuss their model at the end of Section 3.1. In this paper, we expand upon Bell and Keeney (2009) to incorporate a more general interpretation of altruism, nonlinear value functions, and more than two individuals.

Bergstrom (1999) analyzes the interdependencies of individuals' altruistic preferences in an intergenerational setting using an approach similar to our models for individuals. He presents a result that is analogous to a result from our second model ($n > 2$) that we reach via the implicit function theorem. Bergstrom's approach imposes an additional requirement called *coherence*, which is presented and discussed by Pearce (2008). We do not require coherence directly; rather, we will argue that it is straightforward for a decision analyst to elicit coherent altruistic value functions from a decision maker. Further details are provided in Sections 3.1, 3.2, and 4. We also expand the result to a group setting, which allows for a more easily applicable condition based on the implicit function theorem.

Our paper develops models of *pure* altruism, as does the aforementioned work of Bergstrom (1999), Bell and Keeney (2009), and Simon (2016). Pure altruism is distinct from value derived from the *act* of contributing to another's outcome or well-being. This dichotomy is stated explicitly by Arrow (1975) and many subsequent papers including, e.g., Andreoni (1990), Ribar and Wilhelm (2002), McCardle et al. (2009), and Ottoni-Wilhelm et al. (2017). While it is certainly possible that a purely altruistic individual could apply the value functions in this paper to charitable donations, they are not intended as descriptive models of that setting.

We develop a model in Section 3.3 that involves aggregation of individual values through the use of group preferences. Some prior papers also examine an individual's interaction with the overall group rather than with other individuals. Chen and Plott (2002) discuss the aggregation of individual beliefs into a single group belief. This idea is also analyzed by Forsythe et al. (1995) and Forsythe et al. (1999). Nord et al. (1999) study similar issues in the context of health care, evaluating different health programs by assessing individuals' preferences over the health outcomes of others, with an emphasis placed on equity.

Any altruistic preference model depends, whether implicitly or explicitly, on some form of interpersonal comparison of value or utility, which some will argue is impossible to achieve. This issue, which is central to welfare economics, has been discussed by, e.g., Robbins (1938), Kaldor (1939), Harsanyi (1955), Hammond (1976), and Binmore (2009). From a social decision maker's perspective, it is not clear that meaningful tradeoffs can be made between the value or utility of two (or more) different individuals. In this paper, we rely on interpersonal comparison of value only within an individual's preferences; that is, we do not attempt to determine optimal social outcomes. Altruism is, in fact, a specific type of externality. Consequently, most of the ideas and results based on altruism in this paper can be applied with minimal adjustments to externalities of preferences in general.

We frame altruistic preferences as characteristics of individual decision makers that can be captured for prescriptive purposes. However, given the elusiveness and controversy surrounding this topic, it is important to note that there is a large stream of literature using evolutionary models to explain and justify the existence of altruism in rational decision makers. To give just a few examples, Trivers (1971), Alexander (1987), Simon (1990), Bergstrom and Stark (1993), and Nowak and Sigmund (2005) all provide compelling motivation to understand and model altruistic preferences from a decision analysis perspective.

3. Model and Theoretical Results

Contrary to the approach taken by many researchers in the past, we aim to analyze altruistic preference models in a normative fashion, with the goal of eventually using the models in prescriptive settings. In particular, we are interested in preference structures over outcomes affecting multiple altruistic individuals. This section is divided into three subsections, corresponding to the cases of two individuals, n individuals ($n > 2$), and a group. Each subsection provides existence results for altruistic value functions. These value functions could serve as inputs to a wide range of decision settings, including individual decisions, group decisions, bargaining models, or optimal social planning.

3.1. Two Individuals

Consider a decision made by Person 1, yielding outcome x_1 for the decision maker, for which an outcome x_2 is also experienced by Person 2. For example, suppose Person 1 orders pizza for dinner for both individuals, or decides how large a gift to give to Person 2. Let X denote the set of possible outcomes to Person 1 and Person 2, where $x \in X$ can be expressed as $x = (x_1, x_2)$ with x_1 and x_2 within closed intervals of the real line. The outcomes are expressed as vectors to allow for the possibility that the two people experience different outcomes; it is straightforward to restrict X to outcomes in which $x_1 = x_2$ when appropriate.

There are two potential preference relations that Person 1 might express over X. The first, denoted as $\succsim_1$, is a preference relation over X that ignores x_2, i.e. effectively a preference relation over X_1: the set of possible levels of x_1. The second, denoted as $\succsim_{1'}$, is a preference relation over X that takes into account both x_1 and x_2. Similarly, Person 2 might express a preference relation $\succsim_2$ over X that ignores x_1, and a preference relation $\succsim_{2'}$ that takes into account x_1 and x_2. In a two-person setting, Simon (2016) defines an altruistic preference relation as one in which a Pareto superior outcome is always preferred; specifically, holding one person's outcome fixed, a change in the other person's outcome that (s)he considers an improvement must lead to a more preferred outcome. We adopt that definition here. Note that this assumption is not always made in settings with multiple individuals; for instance, any model that includes social comparisons (e.g. Bolton (1991)) implies that an individual would be made better off when others suffer (though the models generally are not framed that way). A model that includes equity also involves direct comparisons between individuals, and may have this implication as well.

Simon (2016) shows that under relatively weak conditions, altruistic preferences can be represented by an ordinal altruistic value function $V(x_1, x_2)$. (The conditions are weaker versions of transitivity and continuity, and when restricted to altruistic preferences only, they imply transitivity, continuity, and completeness.) These conditions also imply that V is *decomposable* (Krantz et al. 1971), in that it can be expressed as $V(x_1, x_2) = f(v_1(x_1), v_2(x_2))$, where v_1 and v_2 can be interpreted as single-attribute value functions over X_1 and X_2, respectively. It is straightforward to observe that if Person 1 and Person 2 are both altruistic and the relevant preference conditions are met, their altruistic preferences can thus be represented by $V_1(x_1, x_2) = f_1(v_{11}(x_1), v_{12}(x_2))$ and $V_2(x_1, x_2) = f_2(v_{21}(x_1), v_{22}(x_2))$, respectively, where v_{ij} is Person i's single-attribute value function over X_j.

If the Thomsen condition or hexagon condition (Karni and Safra 1998) holds for each individual, then the altruistic value functions have the additive form $V_1(x_1, x_2) = w_{11}v_{11}(x_1) + w_{12}v_{12}(x_2)$ and $V_2(x_1, x_2) = w_{21}v_{21}(x_1) + w_{22}v_{22}(x_2)$, where w_{ij} is a weight placed by Person i on the single-attribute value function over X_j, and each single-attribute value function is unique up to positive linear transformations. These conditions assert that preferences over x_1 and x_2 are additively separable; they are cancellation conditions implying that when particular indifference judgments are obtained from a decision maker, the tradeoffs associated with those judgments can be concatenated to yield other pairs of outcomes between which the decision maker is indifferent.

However, when we consider a single decision setting that affects both x_1 and x_2, regardless of who is making the decision or how they do so, there is no guarantee that the two single-attribute value functions over X_1 (v_{11} and v_{21}) are related or similar in any way, nor are those over X_2 (v_{12} and v_{22}). When the altruistic preferences of both individuals are of interest, subsequent analysis and assessment will be eased greatly if such a relationship can be established. That is, we would like Person 1's preferences over X_2 to satisfy some type of consistency with Person 2's preferences over X_2, and similarly for X_1. The particular condition(s) required will vary depending on the form of V_1 and V_2. In this paper, we present representation theorems for additive altruistic value functions. Analogous theorems for other functional forms could certainly be developed from other sets of preference conditions. To establish relationships between v_{11} and v_{21}, and between v_{12} and v_{22}, we introduce the following condition:

- Two altruistic individuals' preferences are *midvalue consistent* if for any pair of individual outcomes (x_1^A, x_1^B) for which a tradeoff midvalue exists for both individuals, if x_1^M is a tradeoff midvalue for one individual, then it is also a tradeoff midvalue for the other individual (and similarly for any pair (x_2^A, x_2^B)).

Intuitively, midvalue consistency asserts that if one individual considers the changes from x_i^A to x_i^M and from x_i^M to x_i^B to be "equal" improvements, then the other individual does so as well; see Harvey (1995) for more detail on tradeoff midvalues.

Theorem 1. *Let $\succsim_{1'}$ and $\succsim_{2'}$ satisfy the conditions needed for additive altruistic value functions to exist. There exist single-attribute value functions v_1 and v_2, unique up to positive linear transformations, such that $V_1(x_1, x_2) = w_{11}v_1(x_1) + w_{12}v_2(x_2)$ and $V_2(x_1, x_2) = w_{21}v_1(x_1) + w_{22}v_2(x_2)$ if and only if midvalue consistency holds.*

A proof of Theorem 1 is given in the Appendix. We will refer to v_1 and v_2 as *traditional* value functions throughout the paper.

Thus, if the relevant preference conditions are satisfied, we can express Person 1's altruistic value function more generally as:

$$V_1(x) = f(v_1(x_1), v_2(x_2)), \tag{1}$$

and similarly for Person 2. We refer to this representation as *specific altruism*. We claim, however, that for prescriptive purposes, the following expression would be even more helpful:

$$V_1(x) = f(v_1(x_1), V_2(x)). \tag{2}$$

We refer to Equation 2 as *general altruism*. Both expressions reflect non-paternalistic altruism. (Under paternalistic altruism, an individual has preferences directly over the other individual's outcome, ignoring the other individual's preferences.) For our purposes, we adopt Equation 2 as providing a more accurate representation of altruistic value. Preferences over others' traditional value could be viewed simply as a proxy for preferences over the well-being of others, without relying on such a rigid restriction on how that well-being is derived. That is, one could argue that Equation 2 is more consistent with the spirit of non-paternalistic altruism than Equation 1.

In addition, elicitation of general altruistic preferences is rather straightforward, as discussed in Section 4.

The mathematical challenge of general altruistic value functions is that they involve complex interdependencies between the values of the individuals, making it very difficult to develop a representation theorem that justifies their use. The major concern is that, given assessed general altruistic value functions, it might not be possible to resolve the interdependencies and provide corresponding specific altruistic value functions. That is, it is not guaranteed that the value functions given in Equation 2 can be expressed in the form of the value functions given in Equation 1, which represent the two individuals' altruistic preferences. In this paper, we prove existence results showing that, by meeting a few basic conditions, it is indeed possible to express general altruism only in terms of $v_1(x_1)$ and $v_2(x_2)$, so that $V_1(x)$ and $V_2(x)$ do not contain one another as arguments. Paired with the representation theorem provided for specific altruistic value functions, this allows for the use of general altruistic value functions to represent altruistic preferences. We will begin with a very simple case using an additive form, and then expand to more general structures.

Consider additive general altruistic value functions of the form:

$$\begin{aligned} V_1(x) &= \alpha_1 v_1(x_1) + (1-\alpha_1)V_2(x) \\ V_2(x) &= \alpha_2 v_2(x_2) + (1-\alpha_2)V_1(x) \end{aligned} \tag{3}$$

with $0 \leq \alpha_1, \alpha_2 \leq 1$. The first natural question is to express $V_1(x)$ and $V_2(x)$ only in terms of $v_1(x_1)$ and $v_2(x_2)$; according to Theorem 1, such a form can be used to represent altruistic preference relations. This can be done provided that α_1 and α_2 are not both equal to zero. (The reason we cannot have $\alpha_1 = \alpha_2 = 0$ is that it reduces the expressions to $V_1(x) = V_2(x)$ and $V_2(x) = V_1(x)$, which provides no information about the levels of value actually achieved.) More generally, as well understood, solving a linear system of n equations with $m > n$ unknowns requires the equations to satisfy certain algebraic conditions. The result is the following pair of additive altruistic value functions:

$$\begin{aligned} V_1(x) &= \frac{\alpha_1}{1-(1-\alpha_1)(1-\alpha_2)} v_1(x_1) + \frac{(1-\alpha_1)\alpha_2}{1-(1-\alpha_1)(1-\alpha_2)} v_2(x_2) \\ V_2(x) &= \frac{\alpha_2}{1-(1-\alpha_1)(1-\alpha_2)} v_2(x_2) + \frac{(1-\alpha_2)\alpha_1}{1-(1-\alpha_1)(1-\alpha_2)} v_1(x_1). \end{aligned} \tag{4}$$

The weights capture the relative amounts of overall value derived from one's own traditional value function and the other individual's traditional value function. Note that this result includes the indirect effects that arise from both individuals being altruistic.

The weights in the general altruistic value functions in Equation 3 are required to sum to 1. Though this is often done simply by convention, it has an additional benefit in this setting: it ensures coherence. Without such a restriction, it is possible that any positive initial levels of V_1 and V_2 would lead to both increasing indefinitely, and the only possible solutions would be negative, with V_1 and V_2 decreasing in v_1 and v_2 (Pearce 2008, Bergstrom 1989). (For instance, consider the functions $V_1(x) = v_1(x_1) + 2V_2(x)$ and $V_2(x) = v_2(x_2) + 2V_1(x)$.) Analogous restrictions would need to be imposed on any non-additive form to ensure coherence.

If Person 1 and Person 2 both display general altruism, then both can be made better off as a result of each other's altruism. Bell and Keeney (2009) use the example of a couple having a meal together. If Person 2 enjoys her meal, this makes Person 1 happier, which in turn makes Person 2 happier, which then makes Person 1 happier, ad infinitum. A natural reaction to this realization is that this infinite regress might mean that two people exhibiting general altruism should converge to identical value functions. In fact, this is not the case. This can be checked by assuming that $\alpha_1, \alpha_2 > 0$ and then setting $V_1(x) = V_2(x)$. It turns out that this implies $v_1(x_1) = v_2(x_2)$ for all possible outcomes, which is highly implausible. Thus, incorporating both individuals' altruism does not preclude the possibility that, given two outcomes x^A and x^B, one will prefer x^A and the other will prefer x^B. This conclusion appears to be contrary to the sense of altruism; it implies that even strongly altruistic individuals may prefer alternatives that decrease the overall value of others. However, this is undeniably true for the additive model: If Person 1 and Person 2 ever differ in their traditional preferences over outcomes, then they will never have identical altruistic value functions[1].

[1] There is prior work examining conditions under which individual assessments of social preferences will converge to a common function. For instance, Lehrer (1978) examines an additive model in which each person assigns a set of

This result shows that, excluding a degenerate case, additive general altruistic value functions can be expressed as specific altruistic value functions (for which a representation theorem exists). The next question is: can we expand this result to two-person altruistic preferences without relying on a specific structure? The desired approach is to express these general value functions as:

$$\begin{aligned} V_1(x) &= f_1(v_1(x_1), V_2(x)) \\ V_2(x) &= f_2(v_2(x_2), V_1(x)), \end{aligned} \tag{5}$$

and then express $V_1(x)$ and $V_2(x)$ as functions that do not contain one another as arguments. If we cannot do this for a given x, this means either that there is no possible pair of altruistic values that solves these two expressions for that outcome, or that there is no local unique functional representation of V_1 and V_2. The former implies that the altruistic value functions of the two individuals are in a sense "incompatible" at x, meaning it would be impossible for the individuals to judge this particular outcome using this approach. (This is unlikely to arise in practice; it cannot occur for any commonly used form of value function.) The latter implies that, at x, there are multiple possible functions for V_1 and V_2, suggesting that they are under-specified. A common demonstrative example of this phenomenon is, given the relationship $x^2 + y^2 = 1$, trying to express y as a function of x. It is clear that the points $(1,0)$ and $(-1,0)$ are solutions to the equation, but there is no single functional form for y that applies within an open set containing either of them. In the context of value functions, this presents a challenge, and we will explore this case in more detail.

The assumptions required to resolve the altruistic value interdependencies are now somewhat more involved; they come from the implicit function theorem (IFT), which generalizes the algebraic conditions that were needed in the case of Equation 3. The IFT states that for a continuously differentiable function $F(v) = (f_1(v_1, \ldots, v_{n+m}), \ldots, f_n(v_1, \ldots, v_{n+m})) : \mathbb{R}^{n+m} \to \mathbb{R}^n$ (with variables $v_1, \ldots, v_n, v_{n+1}, \ldots, v_{n+m}$), if the following two conditions hold at a given point $v_0 = (v_{0,n}, v_{0,m})$:

weights to each member of a group. If weights are updated iteratively, all individuals will converge to a common set of weights provided every individual positively influences every other individual (directly or indirectly). However, these are social preferences, not individual preferences.

- $F(v_0) = 0$
- $\det \begin{bmatrix} \frac{\partial f_1}{\partial v_1} & \cdots & \frac{\partial f_1}{\partial v_n} \\ \vdots & \ddots & \\ \frac{\partial f_n}{\partial v_1} & & \frac{\partial f_n}{\partial v_n} \end{bmatrix} \neq 0$, evaluated at $v = v_0$,

then there exists a continuous function $H : \mathbb{R}^m \to \mathbb{R}^n$ defined over a neighborhood about $v_{0,m}$ such that $H(v_{0,m}) = (v_{0,n})$ and $F(H(v_m), v_m) = 0$. That is, for a point that solves the initial system of equations, we can locally express the first n variables as a function of the last m variables, provided the matrix of partial derivatives of f with respect to the first n variables has full rank at the specified point. The matrix of partial derivatives provides a local linear approximation of the system of equations; a determinant of zero is analogous to $\alpha_1 = \alpha_2 = 0$ in the additive case given by Equation 3. For more details on this theorem, see Krantz and Parks (2002). For our purposes, if the first n variables represent the general altruistic values (and the last m variables represent the traditional values), the IFT assures that if both conditions hold over a connected set of points (outcomes), then we have a unique functional representation for the general altruistic values on this set. This is extremely valuable for prescriptive purposes, as it implies that general altruistic value functions can be assessed.

In the two-person case, the goal is to express $V_1(x)$ and $V_2(x)$ as functions of the traditional values $v_1(x_1)$ and $v_2(x_2)$. (For notational simplicity in conditions and proofs, we will write $V_1(x), V_2(x), v_1(x_1)$, and $v_2(x_2)$ as V_1, V_2, v_1, and v_2, respectively, with the understanding that they depend on their associated outcomes.) If a solution exists for Equation 5 at a given x, then this goal turns out to be possible provided that:

$$\frac{\partial f_1}{\partial V_2} \frac{\partial f_2}{\partial V_1} \neq 1, \tag{6}$$

as evaluated at this point, and, of course, that these partial derivatives exist. See the Appendix for the derivation of (6). It should be apparent that this is not a restrictive condition for a given solution; even a small perturbation to the altruistic value functions would resolve a violation of it. In other words, it is possible to compute functions for V_1 and V_2 (at least locally) for a wide range of general altruistic value forms.

Bell and Keeney (2009) raise an objection to general altruism that is a legitimate concern: it leads to "double counting" of value. That is, Person 1 is made better off not only by an outcome that he enjoys, but also by the positive effect that his resulting well-being has on Person 2. When both individuals exhibit general altruism, this leads to an infinite mathematical process for calculating the two values. However, the implicit function theorem approach avoids this issue by solving for the representation in a region at which the interdependencies are satisfied, and can be viewed in this setting as a fixed point theorem. While not a perfect analogy, this is similar to the idea given by Nash (1951) to solve directly for an equilibrium in a game theory model, thus sidestepping a more complex analysis of a sequence of best responses.

The issue of coherence raised by Pearce (2008) could be resolved by asserting further that the left-hand side of (6) cannot be greater than 1, as explored further in the following subsection. It is unlikely that such altruistic value functions would be elicited in practice; for instance, in the case of additive value functions, it would require that at least one of the functions includes a weight greater than 1.

3.2. More Than Two Individuals

The next step is to expand the model to more than two altruistic value functions. We denote the set of outcomes affecting Person i as X_i, Person i's preferences over X_i as $\succsim_i$, Person i's altruistic preferences over X as $\succsim_{i'}$, and Person i's traditional and altruistic value functions as v_i and V_i, respectively. We adapt the Pareto definition of altruistic preferences accordingly; given two outcomes x^A and x^B, if $x_i^A \succsim_i x_i^B$ for all i, then $x^A \succsim_{i'} x^B$ for all i, and if the strict relation holds for at least one pair x_i^A and x_i^B, then $x^A \succ_{i'} x^B$ for all i.

LEMMA 1. *For any individual i, there exists a continuous real-valued function $V_i(x_1, \ldots, x_n)$ such that $V_i(x_1^A, \ldots, x_n^A) \geq V_i(x_1^B, \ldots, x_n^B)$ iff $(x_1^A, \ldots, x_n^A) \succsim_{i'} (x_1^B, \ldots, x_n^B)$ if and only if $\succsim_{i'}$ satisfies the conditions of Simon (2016) adapted to n individuals (detailed conditions are provided in the proof).*

The proof is given in the Appendix. The conditions needed to satisfy Lemma 1 are weaker than those typically needed for a multiattribute value function; this is possible due to the restriction that the preference relation must be altruistic. As in the case of two individuals, these conditions imply that V_i is decomposable, and can be expressed as $V_i(x_1,\ldots,x_n)=f(v_1(x_1),\ldots,v_n(x_n))$, where $v_1,\ldots,v_n$ can be interpreted as traditional value functions. As previously for the case where $n=2$, we will provide a representation theorem for additive altruistic value functions for illustrative purposes; it is certainly possible that analogous representation theorems could be developed for other functional forms.

When $n=2$, if the Thomsen condition or hexagon condition is satisfied, then the altruistic value function will have an additive form. For $n>2$, we instead use the preferential independence condition (Debreu 1960), which in this context states that tradeoffs an individual is willing to make between a subset of individuals' outcomes do not depend on the common set of outcomes experienced by the other individuals. Note that preferential independence of individuals' outcomes precludes explicit considerations of equity or social comparison. If this condition is satisfied, then Person i's altruistic value function has the form:

$$V_i(x_1,\ldots,x_n)=\sum_{j=1}^{n} w_{ij}v_{ij}(x_j) \tag{7}$$

The concept of midvalue consistency extends naturally to n individuals, and allows us to state the following Theorem:

Theorem 2. *Let $\succsim_{1'},\ldots,\succsim_{n'}$ satisfy the conditions needed for additive altruistic value functions $V_1,\ldots,V_n$ to exist. There exist single-attribute value functions $v_1,\ldots,v_n$, unique up to positive linear transformations, such that $V_i(x_1,\ldots,x_n)=\sum_{j=1}^{n} w_{ij}v_j(x_j)$ for all i if and only if midvalue consistency holds across $\succsim_{1'},\ldots,\succsim_{n'}$.*

See the Appendix for the proof of Theorem 2.

For $n>2$, we can again apply the IFT to obtain the conditions required to be able to express each individual's altruistic value function as $V_i(x)=f_i(v_i(x_i),V_1(x),\ldots,V_{i-1}(x),V_{i+1}(x),\ldots,V_n(x))$.

With n individuals, the main condition is somewhat more abstract. At a given solution point, the following determinant must be non-zero:

$$\det \begin{bmatrix} 1 & -\frac{\partial f_1}{\partial V_2} & \cdots & -\frac{\partial f_1}{\partial V_{n-1}} & -\frac{\partial f_1}{\partial V_n} \\ -\frac{\partial f_2}{\partial V_1} & 1 & & & \\ \vdots & & \ddots & & \\ -\frac{\partial f_{n-1}}{\partial V_1} & & & 1 & -\frac{\partial f_{n-1}}{\partial V_n} \\ -\frac{\partial f_n}{\partial V_1} & & & -\frac{\partial f_n}{\partial V_{n-1}} & 1 \end{bmatrix} \neq 0 \qquad (8)$$

(or equivalently that this matrix has full rank or is invertible), and again, that these partial derivatives all exist. See the Appendix for details (under "Derivation of (8)"). We will refer to the matrix in (8) as D. Note that in the case of additive value functions, all of the partial derivatives are constant, and reflect the weight that an individual places on another individual's general altruistic value.

The interpretation of (8) is similar to that for the two-person model. If the set of general altruistic value functions can be solved at a given point where D does not have full rank, a unique representation of V_i over a neighborhood about this point is not guaranteed. It is still possible to compute values for the V_i; D having lower rank would merely imply that they are under-specified. If, on the other hand, D has full rank, then there is a unique functional representation for the V_i, and it is possible to (locally) solve the n-person general altruistic model and express $V_1, \ldots, V_n$ in terms of $v_1, \ldots, v_n$.

Bergstrom (1999) states a similar result in the context of intergenerational preferences. That result also includes a restriction that imposes coherence (Pearce 2008) on the set of value functions, which would be tantamount to requiring that all of the principal minors in the matrix in (8) are positive. In economic input-output models, it is known as the Hawkins-Simon condition (Hawkins and Simon 1949). In this context of this paper, such a restriction is generally not necessary, because violations of it can be avoided when the general altruistic value functions are elicited from the decision maker, as discussed in Section 4. For commonly used forms of value functions, the conventional weighting and scaling approaches used by decision analysts ensure coherence.

The condition given by (8) is a promising, but still somewhat vague result, in that it cannot be easily reduced to a simple intuitive expression of the partial derivatives of f. However, of definite interest are the types of cases where it is, or is not, satisfied. For the determinant to be of lower rank, at least one row (or column) can be expressed as a linear combination of the others. When many of the partial derivatives in (8) are equal to zero, it can be possible to construct simpler expressions that guarantee existence of altruistic value functions. This has an interesting implication: it suggests that an individual's value may be affected by another individual's value, even if the two never directly interact. This is a provocative trait of altruistic models that does not arise in the two-person case, and is the motivation for the approach used in the following subsection.

3.3. Group Preferences

We will now discuss an alternate formulation for n individuals that fits very well into the general altruism framework, and has a simpler required condition while still capturing the altruistic implications discussed previously. We create an artificial "group entity" as the $n+1^{\text{st}}$ person, and let each of the n individuals display altruism only toward this group entity. That is, the individuals are concerned with the success of the group, but not directly with the values of the other individuals. It is similar in motive to the combination of self and group preferences used by Margolis (1984), and Hamilton (1963) gives an evolutionary justification for preferences of this type. The group's fundamental performance level can be defined and measured, and will be considered as one element of any outcome x. We assert that the overall success or well-being of the group, however, is also permitted to depend (partially) on the well-being of each individual. The group formulation imposes a constraint on the preferences of the n individuals. The benefit, however, is that it will lead to altruistic value functions that are far easier to assess.

As an example, consider a volunteer social action group deciding on their next activity. Each individual will have a higher level of value if the activity goes well. However, the success of the group depends both on the success of the activity in advancing the group's cause and on the individuals being happy and motivated to dedicate their time and energy to it. Thus, we have the same type of interrelated value dynamic that arose in the previous models. This could be valuable when considering how people in individualistic vs. collectivistic cultures trade off individual and group outcomes.

Let $x_g \in X_g$ denote a *group outcome*. We can then express an outcome $x \in X$ as $x_1, \ldots, x_n, x_g$. As previously, to let X be as general as possible; we allow any combination of $x_1, \ldots, x_n, x_g$, recognizing that in reality, x_g might be determined (or at least constrained) by $x_1, \ldots, x_n$. Let $\succsim_g$ denote a preference relation over X_g, and $\succsim_{g'}$ denote a preference relation over X. We can define altruism for $\succsim_{g'}$ as previously; the group entity will always prefer Pareto superior outcomes (with x_g included as part of the outcomes). The representation theorem given previously for $V_i(x)$ applies to $V_g(x)$ as well, and we will again explore the additive case.

If $\succsim_{g'}$ satisfies preferential independence, then:

$$V_g(x) = w_{gg} v_{gg}(x_g) + \sum_{j=1}^{n} w_{gj} v_{gj}(x_j), \tag{9}$$

and an adapted midvalue consistency condition can be used as previously to establish that the single-attribute value functions for any specific individual are positive linear transformations of one another, such that:

$$\begin{aligned} V_i(x) &= \left[\sum_{j=1}^{n} w_{ij} v_j(x_j)\right] + w_{ig} v_g(x_g) \\ V_g(x) &= \left[\sum_{j=1}^{n} w_{gj} v_j(x_j)\right] + w_{gg} v_g(x_g). \end{aligned} \tag{10}$$

The group entity is, to this point, simply a differently-labeled $n+1^{\text{st}}$ individual. However, the following condition will allow for a simpler form of the altrustic value functions in this setting:

- Altruistic individuals' preferences $\succsim_{1'}, \ldots, \succsim_{n'}$ and group preferences $\succsim_{g'}$ that satisfy the conditions needed for altruistic value functions to exist are *group homogeneous* if for any individual i and outcomes (x_i, x^A_{-i}) and (x_i, x^B_{-i}), $(x_i, x^A_{-i}) \succsim_{i'} (x_i, x^B_{-i})$ iff $(x_i, x^A_{-i}) \succsim_{g'} (x_i, x^B_{-i})$.

Group homogeneity asserts that an individual's preferences regarding other individuals' outcomes are determined by the group entity's preference over those outcomes, i.e., the altruistic component of each individual's preferences must adhere to a common ordering of outcomes. A clear implication of group homogeneity is that for any subset of individuals, each is willing to make the same tradeoffs between outcomes of individuals not in that subset. The purpose of this condition is to require that the individuals in the group are altruistic only toward the group as a whole, not toward other individuals directly. A decision analyst should verify that the condition holds before using the approach in this subsection.

Theorem 3. *If $\succsim_{1'}, \ldots, \succsim_{n'}, \succsim_{g'}$ satisfy the conditions needed for additive altruistic value functions to exist, then these altruistic value functions have the form: $V_i(x) = w'_{ii} v_i(x_i) + w'_{ig} V_g(x)$ and $V_g(x) = w'_{gg} v_g(x_g) + \sum_{j=1}^{n} w'_{gj} V_j(x_j)$ iff group homogeneity is satisfied and the applicable IFT condition is met.*

The proof of Theorem 3 is given in the Appendix. Unlike the preceding representation theorems, it requires that an IFT condition is met, as detailed in the proof. This arises because Theorem 3 uses general altruistic value functions. However, the IFT condition is much simpler in the additive case, and in particular does not depend on x. Note also that the weights, in general, are not equal to the weights of the altruistic value functions in (10).

The IFT condition presented in this section concerns general altruistic value functions of the form:

$$\begin{aligned} V_i(x) &= f_i(v_i(x_i), V_g(x)) \\ V_g(x) &= f_g(v_g(x_g), V_1(x), \ldots, V_n(x)). \end{aligned} \tag{11}$$

Theorem 3 establishes conditions for the existence of additive forms of V_i and V_g that satisfy (11); the IFT condition will apply to any functional forms in which no individual's general altruistic value appears as an argument in any other individual's general altruistic value function. The condition required to express V_i and V_g as functions of $v_1, \ldots, v_n$ and v_g is simpler than the condition given in the previous subsection. The matrix that must now have full rank, as specified in the IFT, is much sparser, and the required condition for a local functional representation at a given solution point reduces to:

$$\sum_i \frac{\partial f_i}{\partial V_g} \frac{\partial f_g}{\partial V_i} \neq 1. \tag{12}$$

Details can be found in the Appendix. Intuitively, this condition asserts that the effects of the individuals' values on the group and the group value on the individuals cannot precisely cancel out. This is conceptually similar to condition (6) in the two-person case. In fact, the two-person case is mathematically equivalent to the special case of the group model with one individual (plus the artificial group entity). As long as this condition holds, it is possible to express locally each individual's value and the group value as functions of $v_1(x_1), \ldots, v_n(x_n)$ and $v_g(x_g)$.

4. Assessment

In this section, we briefly discuss techniques that can be used to elicit altruistic preference information from decision makers. We will first present standard approaches that can be adapted to this setting with minimal modification, and will then explore the aspect of assessment unique to general altruism.

The traditional single-attribute value functions can be assessed using any of the techniques commonly used by decision analysts for this purpose. For example, midvalue splitting (Keeney and Raiffa 1976, Kirkwood 1997) can be applied repeatedly to approximate these value functions to the desired degree of precision. Alternatively, the value functions can be fit to a particular functional form, if it is judged that preferences satisfy the conditions that give rise to the form. Note that in the context of altruistic decisions, it would be necessary either to elicit traditional value functions from multiple individuals, or to assume that one individual has an adequate understanding of the other individuals' preferences.

If the general altruistic value functions are additive, then standard weight assessment procedures can be applied with little or no modification as well. For example, the value tradeoff method (Keeney and Raiffa 1976, Eisenführ et al. 2010) can be used to identify tradeoffs between two individuals' values for which the decision maker (or another individual) is indifferent, allowing the analyst to solve for a set of weights. As for the traditional value functions, either elicitation from multiple individuals or an understanding of all individuals' preferences would be necessary.

It should be noted, however, that the single-attribute value functions and weights (or other parameters for non-additive forms) should be constructed such that the set of general altruistic value functions will be coherent. For additive forms, the standard approach among decision analysts of using a set of non-negative weights that sum to 1 is sufficient, and no further restriction is necessary. (A proof of this is straightforward, since all of the relevant partial derivatives are constant.) Multilinear and multiplicative forms have similar conventions regarding weights and scaling constants that will ensure coherence. The general underlying principle is that altruistic value functions should be scaled such that a marginal increase in V_i does not produce a greater increase in the other altruistic values. The exact requirement will depend on the form of the altruistic value functions, but the issue of coherence does not pose a concern for the forms commonly used by decision analysts to aggregate multiple sources of value.

The task unique to general altruism is conducting assessments that involve other individuals' general altruistic values. These general altruistic value functions are undoubtedly opaque to decision makers, and there is no clear natural or proxy measure that could be used. However, the purpose and motivation of modeling altruism this way is that the concept itself is quite accessible; it is simply a measure of an individual's overall level of satisfaction or well-being with regard to the decision, regardless of how it is achieved. This lends itself well to a constructed scale (Keeney and Gregory 2005), where the minimum and maximum levels of the scale reflect the least and greatest overall value, respectively, that this individual can achieve in this decision setting. The analyst can then check with the decision maker to ensure that value over this constructed scale is linear (e.g. via midvalue splitting), and if it is not, a non-linear single-attribute value function can be assessed over it.

With additive general altruistic value functions, the value tradeoff method can be applied using these constructed scales. For example, in the two-individual case, if the outcomes denote the amounts of leisure time for each member of a married couple this weekend, weights could be assessed by asking questions such as: "*Would you be willing to decrease your leisure time from 4 hours to 2 hours if it would change your spouse's overall value from its lowest possible value to its highest possible value?*" Once the appropriate set of indifference judgments is obtained, it is possible to solve for weights as typically done when using the value tradeoff method.

Of course, for non-additive forms of general altruistic value functions, it is not guaranteed that the concept of a "weight" would be meaningful, but this challenge is not unique to the context of altruism.

5. Examples

In this section, we provide two example applications of general altruistic preferences to demonstrate the application of the concepts presented and the types of insights that can be drawn. The first application is a two-person model, and the second is a group model.

5.1. Dining Couple

Bell and Keeney (2009) provide an example of a couple going out to eat, and use specific altruistic value functions. We extend this to general altruistic value functions. Consider the following value functions for two individuals going out to dinner:

$$\begin{aligned} V_1(x) &= 0.2v_1(x_1) + 0.8V_2(x), \\ V_2(x) &= 0.5v_2(x_2) + 0.5V_1(x), \end{aligned} \tag{13}$$

where $v_1(x_1) = \sqrt{x_1}$ and $v_2(x_2) = 1 - x_2$. In this example, x is a measure of how spicy the food at the restaurant is ($0 \leq x \leq 1$), where $x_1 = x_2 = x$. Person 1 prefers spicier food (with diminishing returns) but cares much more about Person 2's overall value, while Person 2 prefers less spicy food and places an equal weight on Person 1's overall value. The choices of value functions are entirely for illustrative purposes.

We begin by choosing an arbitrary value of $x : 0 \leq x \leq 1$ to demonstrate the first (and simpler) IFT condition. Take, for instance, $x = 0.5$. A solution exists for this outcome; it is given by:

$$v_1 = \frac{\sqrt{2}}{2},\ v_2 = \frac{1}{2},\ V_1 = \frac{2 + 2\sqrt{2}}{6},\ V_2 = \frac{5 + \sqrt{2}}{12}. \tag{14}$$

(It is straightforward to show that a solution exists for any $x : 0 \leq x \leq 1$.) Using the implicit function theorem as described earlier, the requirement for a local functional representation of V_1 and V_2 is that:

$$\frac{\partial f_1}{\partial V_2}\frac{\partial f_2}{\partial V_1} \neq 1, \tag{15}$$

evaluated at this point. The product of the two partials is 0.4 regardless of the value of x, so it is clear that unique altruistic value functions exist. Based on the weights and the traditional value functions, we can express the general altruistic values as the following functions of the traditional values:

$$V_1 = \frac{1}{3}v_1 + \frac{2}{3}v_2,\ V_2 = \frac{5}{6}v_2 + \frac{1}{6}v_1. \tag{16}$$

The general altruistic values can also be expressed in terms of x, keeping in mind that $x_1 = x_2 = x$:

$$V_1 = \frac{2 - 2x + \sqrt{x}}{3},\ V_2 = \frac{5 - 5x + \sqrt{x}}{6}. \tag{17}$$

The left side of Figure 1 shows the resulting general altruistic value functions, along with the traditional value functions (over x). For illustrative purposes, the general altruistic value functions are rescaled in Figure 1 to have a range of 0-1 over the possible outcomes (and the weights in the functions have been adjusted accordingly). The right side of Figure 1 zooms in on the upper-left corner of the first graph, omits the traditional value functions, and adds shading to show the set of Pareto optimal values of x. For $x < 0.01$, a slight increase in spiciness would be preferable to both people, and for $x > 0.06$, a slight decrease in spiciness would be preferable to both people. These two individuals should undoubtedly choose a restaurant for which spiciness x is low, between 0.01 and 0.06. The most desirable outcome within that range would depend on either their relative bargaining powers, or some type of interpersonal comparison of value outside the scope of this paper. (See, e.g., Keeney (2013).)

Note that we cannot rescale any of the value functions individually without affecting the results. Any permissible transformation (i.e. one that preserves the preferences being represented) would have to be applied in conjunction with corresponding transformations of the other functions. For example, if we were to rescale one of the traditional value functions using a positive linear transformation, the weights in that individual's general altruistic function would have to be adjusted accordingly.

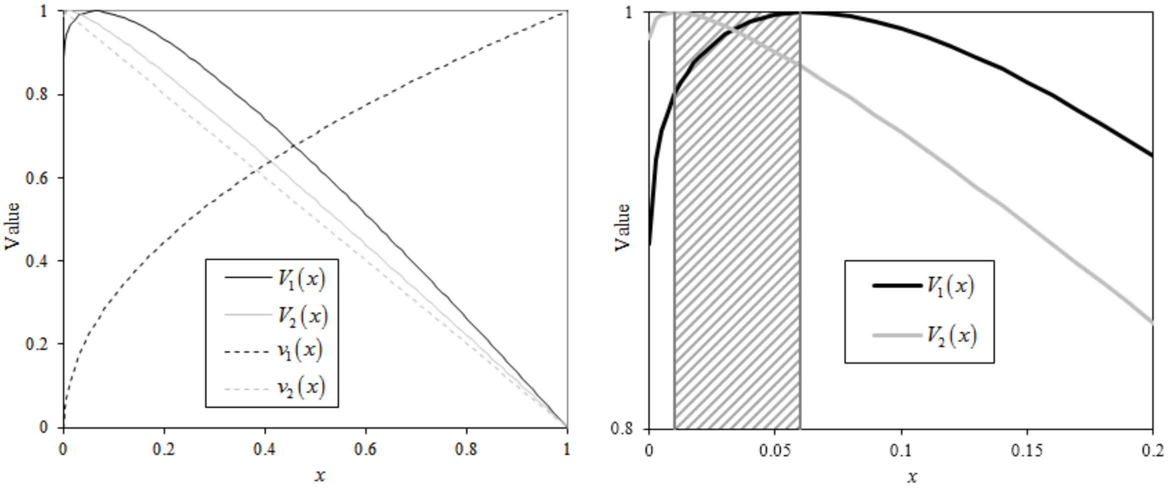

Figure 1 **Dining couple example: value functions (left) and Pareto optimal outcomes $x = 0.01$ to $x = 0.06$ (right)**

5.2. Volunteers in a Social Action Group

In this example, we model a volunteer social action group. There are eight individuals, each of whom has the value function:

$$V_i(x) = 0.75v_i(x_i) + 0.25V_g(x), \tag{18}$$

with $v_i(x_i) = 1 - x_i$. The group value function is given by:

$$V_g(x) = 0.6v_g(x_g) + \sum_{i=1}^{8} 0.05V_i(x), \tag{19}$$

with $v_g(x_g) = \sqrt{x_g}$. In this context, x_i represents the proportion of waking hours that each individual spends promoting the group's next activity ($0 \leq x_i \leq 1$); let x_g be the average value of x_i. In isolation, each individual prefers to spend as little time as possible doing this, but this is counteracted somewhat by the positive effect that it has on the group's success. For simplicity, we allow the individuals to consider only "fair" outcomes, in which each person devotes the same proportion of time (i.e. x_i is equal for all i, and $x_g = x_i$). The group's success is determined mostly by the proportion of time the individuals spend doing promotion (with diminishing returns), but also by the individuals' level of overall well-being.

As in the two-individual example, we begin by arbitrarily choosing $x_i = 0.5$ to demonstrate the first IFT condition. A solution exists for this outcome, given by:

$$v_i = \frac{1}{2},\ V_i = \frac{\sqrt{2}+5}{12} \text{ for } i = 1, \ldots, 8,\ v_g = \frac{\sqrt{2}}{2},\ V_g = \frac{2\sqrt{2}+1}{6}. \tag{20}$$

A local function representation for the V_i and V_g requires that the IFT condition in (12) is satisfied at this point. It is straightforward to compute:

$$\frac{\partial f_i}{\partial V_g}\frac{\partial f_g}{\partial V_i} = (0.25)(0.05) = 0.0125. \tag{21}$$

Since there are eight individuals,

$$\sum_i \frac{\partial f_i}{\partial V_g}\frac{\partial f_g}{\partial V_i} = (8)(0.0125) = 0.1 \neq 1. \tag{22}$$

Thus, the IFT condition will not be violated for any outcomes and, analogous to the two-individual case, functional representations for V_i and V_g apply over the range of possible outcomes. The general altruistic value functions given initially can be expressed as:

$$V_i(x) = \frac{5v_i(x_i) + v_g(x_i)}{6} \tag{23}$$

and:

$$V_g(x) = \frac{16v_g(x_i) + \sum_{i=1}^{8} v_i(x_i)}{24}. \tag{24}$$

The resulting altruistic and traditional value functions (over x_i) are shown in Figure 2. As in the two-individual example, V_i and V_g are rescaled to have a range of 0-1 over the possible outcomes for illustrative purposes. The group's traditional and altruistic values increase with x_i as individuals spend more time promoting. From the altruistic perspective of the individuals in the group, the most desirable choice of x_i is $x_i^* = 0.01$, as indicated by the peak of the solid dark line in Figure 2 (equivalent to approximately 1.16 hours per week for an individual sleeping eight hours per night). If they were not altruistic, their most desirable choice would be zero, since their traditional values decrease monotonically as they spend more time promoting. The effect of including altruism is that a small amount of promotion by the individuals is Pareto superior to none; it is preferred by all individuals, and improves the well-being of the group.

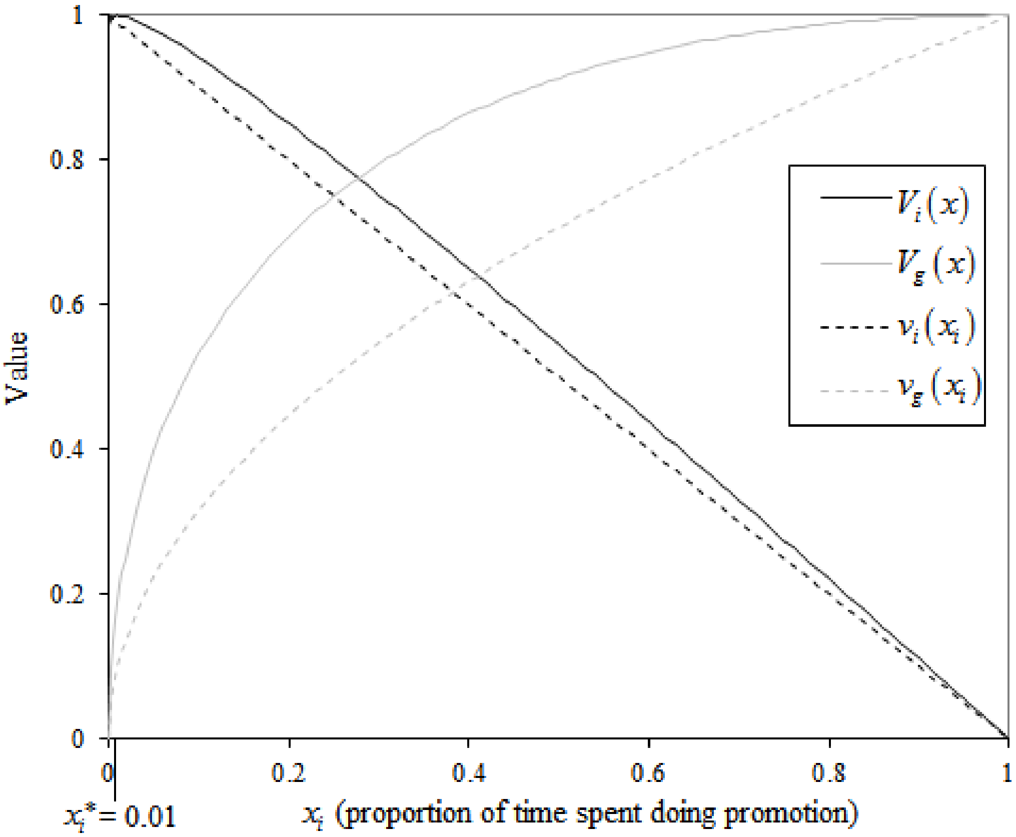

Figure 2 **Social action group example: the individual and group value functions**

5.3. Discussion of Additional Examples

The two previous examples were chosen for simplicity and ease of exposition. There are many other types of decisions for which this approach is applicable.

For instance, Bergstrom (2006) explores an example of a couple with (additive) altruistic preferences choosing between two apartments. When the model of altruistic preferences includes altruism only with respect to the individuals' values for the apartment itself, they can justify paying for the larger apartment despite its cost exceeding the sum of their individual valuations. However, when their altruistic preferences are expanded to include the individuals' values over money as well, the larger apartment can be justified if and only if the sum of their individual gains in valuation exceeds the additional cost. This result is then expanded to a broader analysis of public and private goods.

Game theory settings can also incorporate general altruistic value functions. Ray and Vohra (in press) find that in such games, every Nash equilibrium is a Pareto optimal solution provided that basic coherence conditions are satisfied. Several example are provided, including firms' decisions of whether or not to industrialize, and how introducing a public regulator affects the model. (In the former case, the firms are not explicitly altruistic; however, their payoffs all depend on national income, which is affected by all of their payoffs, generating the same type of interdependent externalities.)

Though the current paper provides representation theorems for additive specific altruistic value functions only, the application of the approach is similar for non-additive functions. For instance, if we were to modify the dining couple example from Section 5.1 as follows, with Person 2 having a non-additive general altruistic value function:

$$\begin{aligned} V_1(x) &= 0.2v_1(x_1) + 0.8V_2(x), \\ V_2(x) &= v_2(x_2)V_1(x), \end{aligned} \tag{25}$$

with v_1 and v_2 unchanged and $x_1 = x_2 = x$ as previously, the corresponding specific altruistic value functions become:

$$\begin{aligned} V_1(x) &= \frac{0.2v_1(x_1)}{1 - 0.8v_2(x_2)}, \\ V_2(x) &= \frac{0.2v_1(x_1)v_2(x_2)}{1 - 0.8v_2(x_2)}, \end{aligned} \tag{26}$$

resulting in a Pareto optimal range of approximately $x \in [0.13, 0.25]$. This range has shifted upward, because choices of x close to zero result in $v_1(x_1)$ being close to zero, which now in turn leads to V_2 being very small. (It is straightforward to confirm that the IFT condition is still satisfied for any choice of x, since $v_2(x_2)$ cannot exceed 1.)

6. Conclusion

Many people incorporate the well-being of others into their own decision making, whether implicitly or explicitly. A major obstacle in modeling and implementing altruistic value functions is the often complicated dynamic of interdependent value that can occur. In this paper, we have examined the general altruistic value concept, where value functions can contain others' altruistic values as arguments, in several different settings.

We provided illustrative representation theorems for the existence of additive altruistic value functions. We began with a two-person model, which had simple conditions for the existence of general altruistic value functions. We expanded this to a general altruistic value model for more than two people, for which we also determined required conditions. We then developed an alternative n-person general altruistic value model using the concept of group preferences, which simplified the required condition, and made clear that it was not overly restrictive.

To resolve the interdependencies imposed by such sets of altruistic value functions, we presented straightforward conditions based on the implicit function theorem. These results, particularly in the second case (more than two people), are similar to results obtained by Bergstrom (1999) examining intergenerational preferences. The current paper is concerned with prescriptive uses of altruistic preferences, and thus omits a condition that is superfluous when the altruistic value functions are developed using standard decision analysis approaches.

In addition, we have provided two illustrative examples of decisions with general altruistic preferences, and guidance on how such preferences might be assessed.

We have shown that, in general, incorporating altruism into preference models is not an analytically insurmountable task. It is nearly always possible to resolve the value interdependencies. We hope that, given these possibility results, further effort can be made in the future to develop effective altruistic decision models.

There are several possible avenues for further research. First, this paper provides representation theorems only for sets of additive altruistic value functions. Representation theorems for other forms would expand the underlying theoretical framework of interdependent altruistic preferences. In addition, all of the theory and examples in this paper use only a single attribute. Representation theorems for altruistic value functions in multi-attribute settings would provide a valuable foundation for a wider range of decision problems.

This paper focuses on formulating and resolving the preferences of multiple altruistic individuals; it would also be useful to explore their impact on the actual decision process. Interdependent altruistic preferences could be incorporated into a group decision analysis approach (Keeney 2013). It is also likely that insights could be gained by incorporating such preferences into cooperative game theory models and bargaining problems.

When individuals are very intensely altruistic and are dividing up a private good, it is possible even with coherent preferences that they will all prefer less of the private good for themselves. This is another compelling class of decision problems to explore further.

References

Alexander, R. D. (1987). *The Biology of Moral Systems.* Aldine de Gruyter, New York.

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

Arrow, K. J. (1975). Gifts and exchanges. In Phelps, E. S., editor, *Altruism, Morality and Economic Theory.* Russell Sage Foundation.

Bell, D. E. and Keeney, R. L. (2009). Altruistic utility functions for joint decisions. In *The Mathematics of Preference, Choice, and Order: Essays in Honor of Peter C. Fishburn.* Springer Berlin, Heidelberg.

Bergstrom, T. (1989). Puzzles: Love and spaghetti, the opportunity cost of virtue. *Journal of Economic Perspectives*, 3(2):165–173.

Bergstrom, T. C. (1999). Systems of benevolent utility functions. *Journal of Public Economic Theory*, 1(1):71–100.

Bergstrom, T. C. (2006). Benefit-cost in a benevolent society. *The American Economic Review*, 96(1):339–351.

Bergstrom, T. C. and Stark, O. (1993). How altruism can prevail in an evolutionary environment. *The American Economic Review*, 83(2):149–155.

Binmore, K. (2009). Interpersonal comparisons of utility. In Kincaid, H. and Ross, D., editors, *Oxford handbook of the philosophy of economic science*, pages 540–559. Oxford University Press, Heidelberg.

Bolton, G. E. (1991). A comparative model of bargaining: Theory and evidence. *The American Economic Review*, pages 1096–1136.

Chen, K.-Y. and Plott, C. R. (2002). Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem. *California Institute of Technology Social Science Working Paper 1131*.

Debreu, G. (1954). Representation of a preference ordering by a numerical function. In Thrall, R. M., Coombs, C. H., and Davis, R. L., editors, *Decision Processes*, pages 159–165. Wiley, New York.

Debreu, G. (1960). Topological methods in cardinal utility theory. In K. J. Arrow, S. Karlin, P. S., editor, *Mathematical Methods in the Social Sciences*, pages 16–26. Stanford University Press, Stanford, CA.

Debreu, G. (1964). Continuity properties of paretian utility. *International Economic Review*, 5(3):285–293.

Eisenführ, F., Weber, M., and Langer, T. (2010). *Rational Decision Making*. Springer-Verlag, Berlin.

Forsythe, R., Frank, M., Krishnamurthy, V., and Ross, T. W. (1995). Using market prices to predict election results: the 1993 ubc election stock market. *Canadian Journal of Economics*, pages 770–793.

Forsythe, R., Rietz, T. A., and Ross, T. W. (1999). Wishes, expectations and actions: a survey on price formation in election stock markets. *Journal of Economic Behavior & Organization*, 39(1):83–110.

Hamilton, W. D. (1963). The evolution of altruistic behavior. *The American Naturalist*, 97(896):354–356.

Hammond, P. (1976). Why ethical measures of inequality need interpersonal comparisons. *Theory and Decision*, 7(4):263–274.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *The Journal of Political Economy*, 63(4):309–321.

Harvey, C. (1995). Proportional discounting of future costs and benefits. *Mathematics of Operations Research*, 20(2):381–399.

Hawkins, D. and Simon, H. A. (1949). Some conditions of macroeconomic stability. *Econometrica*, 17(3/4):245–248.

Kaldor, N. (1939). Welfare propositions of economics and interpersonal comparisons of utility. *The Economic Journal*, 49(195):549–552.

Karni, E. and Safra, Z. (1998). The hexagon condition and additive representation for two dimensions: an algebraic approach. *Journal of Mathematical Psychology*, 42(4):393–399.

Keeney, R. L. (2013). Foundations for group decision analysis. *Decision Analysis*, 10(2):103–120.

Keeney, R. L. and Gregory, R. S. (2005). Selecting attributes to measure the achievement of objectives. *Operations Research*, 53(1):1–11.

Keeney, R. L. and Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* John Wiley and Sons, New York.

Kirkwood, C. (1997). *Strategic Decision Making: Multiobjective Decision Analysis with Spreadsheets.* Duxbury Press, Belmont, CA.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of Measurement, Volume I: Additive and Polynomial Representations.* Academic Press, New York and London.

Krantz, S. G. and Parks, H. R. (2002). *The Implicit Function Theorem: History, Theory, and Applications.* Springer Science & Business Media.

Lehrer, K. (1978). A theory of social rationality. In Hooker, C., Leach, J., and McClennen, E. F., editors, *Foundations and Applications of Decision Theory.* D. Reidel Publishing Company.

Margolis, H. (1984). *Selfishness, altruism, and rationality.* University of Chicago Press.

McCardle, K. F., Rajaram, K., and Tang, C. S. (2009). A decision analysis tool for evaluating fundraising tiers. *Decision Analysis*, 6(1):4–13.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, 54(2):286–295.

Nord, E., Pinto, J. L., Richardson, J., Menzel, P., and Ubel, P. (1999). Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Economics*, 8(1):25–39.

Nowak, M. A. and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437:1291–1298.

Ottoni-Wilhelm, M., Vesterlund, L., and Xie, H. (2017). Why do people give? testing pure and impure altruism. *American Economic Review*, 107(11):3617–3633.

Pearce, D. G. (2008). Nonpaternalistic sympathy and the inefficiency of consistent intertemporal plans. In Jackson, M. O. and McLennan, A., editors, *Foundations in Microeconomic Theory*, pages 213–231. Springer, Berlin.

Ray, D. and Vohra, R. (in press). Games of love and hate. *Journal of Political Economy.*

Ribar, D. C. and Wilhelm, M. (2002). Altruistic and joy-of-giving motivations in charitable behavior. *Journal of Political Economy*, 110(2):425–457.

Robbins, L. (1938). Interpersonal comparisons of utility: A comment. *The Economic Journal*, 48(192):635–641.

Savage, L. (1954). *The Foundations of Statistics.* John Wiley, New York.

Simon, H. A. (1990). A mechanism for social selection and successful altruism. *Science*, 250(4988):1665–1668.

Simon, J. (2016). On the existence of altruistic value and utility functions. *Theory and Decision*, 81(3):371–391.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1):35–57.

7. Appendix

Proof of Theorem 1: Let $\succsim_{1'}$ and $\succsim_{2'}$ satisfy midvalue consistency. We can express the altruistic value functions as $V_1'(x_1,x_2) = w_{11}v_{11}(x_1) + w_{12}v_{12}(x_2)$ and $V_2'(x_1,x_2) = w_{21}v_{21}(x_1) + w_{22}v_{22}(x_2)$. Arbitrarily, let (x_1^A, x_1^B) have a common tradeoff midvalue x_1^M. That means there exist $x_2^{1L}, x_2^{1H}, x_2^{2L}, x_2^{2H}$ such that $(x_1^M, x_2^{1L}) \sim_{1'} (x_1^A, x_2^{1H})$, $(x_1^B, x_2^{1L}) \sim_{1'} (x_1^M, x_2^{1H})$, $(x_1^M, x_2^{2L}) \sim_{2'} (x_1^A, x_2^{2H})$, and $(x_1^B, x_2^{2L}) \sim_{2'} (x_1^M, x_2^{2H})$. Expressing these indifference relations via the altruistic value functions yields:

$$w_{11}v_{11}(x_1^M) + w_{12}v_{12}(x_2^{1L}) = w_{11}v_{11}(x_1^A) + w_{12}v_{12}(x_2^{1H})$$

$$w_{11}v_{11}(x_1^B) + w_{12}v_{12}(x_2^{1L}) = w_{11}v_{11}(x_1^M) + w_{12}v_{12}(x_2^{1H})$$

$$w_{21}v_{21}(x_1^M) + w_{22}v_{22}(x_2^{2L}) = w_{21}v_{21}(x_1^A) + w_{22}v_{22}(x_2^{2H})$$

$$w_{21}v_{21}(x_1^B) + w_{22}v_{22}(x_2^{2L}) = w_{21}v_{21}(x_1^M) + w_{22}v_{22}(x_2^{2H})$$

or, after combining the first two and last two equations and canceling out common terms:

$$v_{11}(x_1^M) = 0.5(v_{11}(x_1^A) + v_{11}(x_1^B))$$

$$v_{21}(x_1^M) = 0.5(v_{21}(x_1^A) + v_{21}(x_1^B)).$$

Because x_1^A and x_1^B were chosen arbitrarily, and continuity of the altruistic preference relations ensures that infinitesimally similar levels will always have a tradeoff midvalue (i.e. that these equalities can be established), it must be true that the functions v_{11} and v_{21} are positive linear transformations of one another. The analogous argument can be made to show that v_{12} and v_{22} are positive linear transformations of one another.

To establish the converse, let tradeoff midvalues x_1^M and $x_1^{M'}$ exist for Person 1 and Person 2, respectively, for arbitrary (x_1^A, x_1^B), and let $v_{21} = cv_{11}$, where $c > 0$. Then:

$$w_{11}v_{11}(x_1^M) + w_{12}v_{12}(x_2^{1L}) = w_{11}v_{11}(x_1^A) + w_{12}v_{12}(x_2^{1H})$$

$$w_{11}v_{11}(x_1^B) + w_{12}v_{12}(x_2^{1L}) = w_{11}v_{11}(x_1^M) + w_{12}v_{12}(x_2^{1H})$$

$$w_{21}cv_{11}(x_1^{M'}) + w_{22}v_{22}(x_2^{2L}) = w_{21}cv_{11}(x_1^A) + w_{22}v_{22}(x_2^{2H})$$

$$w_{21}cv_{11}(x_1^B) + w_{22}v_{22}(x_2^{2L}) = w_{21}cv_{11}(x_1^{M'}) + w_{22}v_{22}(x_2^{2H})$$

After combining and canceling out common terms (note that c disappears), we obtain:

$$v_{11}(x_1^M) = 0.5(v_{11}(x_1^A) + v_{11}(x_1^B))$$

$$v_{11}(x_1^{M'}) = 0.5(v_{11}(x_1^A) + v_{11}(x_1^B)),$$

or $v_{11}(x_1^M) = v_{11}(x_1^{M'})$. Therefore, by definition of v_{11}, $x_1^M \sim_i x_1^{M'}$, and the substitution condition of Simon (2016) ensures that $x_1^{M'}$ is also a tradeoff midvalue for Person 1. (If v_{11} is invertible, then $x_1^{M'} = x_1^M$.) Since the choices of x_1^A and x_1^B were arbitrary, and the analogous argument can be made for arbitrary x_2^A and x_2^B, we can conclude that $\succsim_{1'}$ and $\succsim_{2'}$ are midvalue consistent.

Derivation of (6): In this case, we have four variables (V_1, V_2, v_1, v_2), and would like to express V_1 and V_2 as functions of v_1 and v_2. By moving all terms in (5) to the left-hand side, we obtain:

$$V_1 - f_1(v_1, V_2) = 0$$

$$V_2 - f_2(v_2, V_1) = 0.$$

We state the matrix D of partial derivatives of these two equations with respect to V_1 and V_2:

$$D = \begin{bmatrix} 1 & -\frac{\partial f_1}{\partial V_2} \\ -\frac{\partial f_2}{\partial V_1} & 1 \end{bmatrix}.$$

From the implicit function theorem, we can express V_1 and V_2 as functions of v_1 and v_2 if and only if the determinant of D is non-zero. That is,

$$\det(D) = 1 - \frac{\partial f_1}{\partial V_2}\frac{\partial f_2}{\partial V_1} \neq 0.$$

Thus, the determinant is non-zero provided that:

$$\frac{\partial f_1}{\partial V_2}\frac{\partial f_2}{\partial V_1} \neq 1,$$

which establishes (6).

Proof of Lemma 1: Let the traditional preferences $\succsim_i$ be complete, transitive, and continuous. For the sake of simplicity, we also assume that the traditional preferences are monotonic, with greater values of x_i preferred. Monotonicity is not required, as the attributes can be transformed as needed, but it will greatly reduce the notational burden.

We then restate the conditions of Simon (2016), adapted for n individuals. (Without loss of generality, we will consider the preferences of Person 1 for this proof.)

- $\succsim_{1'}$ satisfies the *substitution* property if for any $x^A, x^B \in X$ such that $x^A \sim_{1'} x^B$, it holds that for any $x^C \in X$, $x^A \succsim_{1'} x^C$ iff $x^B \succsim_{1'} x^C$, and $x^C \succsim_{1'} x^A$ iff $x^C \succsim_{1'} x^B$.
- $\succsim_{1'}$ satisfies the *indifference* property if for any $x^A, x^B, x^C \in X$ such that $x^A \succsim_{1'} x^B$ and $x^B \succsim_{1'} x^C$, there exists a real number $k \in [0,1]$ such that $x^C + km \sim_{1'} x^B$, where $m = (x_1^A - x_1^C, \ldots, x_n^A - x_n^C)$.

The substitution condition states that outcomes between which Person 1 is indifferent may be substituted for one another without affecting the truth of a comparison. The indifference condition is a specific type of solvability; it states that any "intermediate" outcome has an equally preferable outcome that is a convex combination of the more and less preferred outcomes.

To establish that these conditions imply the existence of an altruistic value function, we must show that for altruistic $\succsim_{1'}$, these conditions imply that $\succsim_{1'}$ is complete, transitive, and continuous over X, in which case the results of Debreu (1954, 1964) will apply.

Completeness: Consider $x^A, x^B \in X$. If $x_i^A \succsim_i x_i^B$ for all i, then $x^A \succsim_{1'} x^B$, since x^A is Pareto superior. The same argument applies if x^B is Pareto superior. We will focus on the nontrivial case in which there exist at least one choice of i and j such that $x_i^A \succ_i x_i^B$ and $x_j^B \succ_j x_j^A$.

Consider the outcomes $x^*, x^0 \in X$, where, for all i, $x_i^* = x_i^A$ if $x_i^A \succsim_i x_i^B$, otherwise $x_i^* = x_i^B$, and $x_i^0 = x_i^A$ if $x_i^B \succsim_i x_i^A$, otherwise $x_i^* = x_i^B$. That is, x^* is the vector of more preferred outcomes for all individuals, and x^0 is the vector of less preferred outcomes for all individuals. Clearly, $x^* \succsim_{1'} x^A, x^* \succsim_{1'} x^B, x^A \succsim_{1'} x^0$, and $x^B \succsim_{1'} x^0$.

By the indifference condition, there exist convex combinations $x^{A'}$ and $x^{B'}$ of x^* and x^0 such that $x^{A'} \sim_{1'} x^A$ and $x^{B'} \sim_{1'} x^B$. By definition of x^* and x^0, either $x^{A'}$ is Pareto superior to $x^{B'}$ or $x^{B'}$ is Pareto superior to $x^{A'}$ (or both), which implies that $x^{A'} \succsim_{1'} x^{B'}$, $x^{B'} \succsim_{1'} x^{A'}$, or both. Then, by the substitution condition, it must be true that $x^A \succsim_{1'} x^B$, $x^B \succsim_{1'} x^A$, or both.

Continuity: Consider $x^A, x^B \in X$ such that $x^A \succ_{1'} x^B$, and let x^*, x^0 be defined as previously. The continuity condition used here is that there exists some $\Delta > 0$ such that if $\max_i(|x_i^C - x_i^A|) \leq \Delta$, then $x^C \succ_{1'} x^B$, and if $\max_i(|x_i^C - x_i^B|) \leq \Delta$, then $x^A \succ_{1'} x^C$. (That is, if x^A or x^B is replaced by a distinct but sufficiently similar outcome, the strict preference relation will still hold.) If x^A is Pareto superior to x^B, then such a Δ exists trivially by continuity of $\succsim_i$. If x^A is not Pareto superior to x^B, then it will suffice to show that there exist $x^{A'}, x^{B'} \in X$ such that $x^{A'}$ is indifferent to x^A and Pareto superior to x^B, and $x^{B'}$ is indifferent to x^B and Pareto inferior to x^A. The indifference condition asserts that there must be a convex combination of x^B and x^* to which x^A is indifferent. Let $x^{A'}$ denote that convex combination. By definition of x^*, $x^{A'}$ is Pareto superior to x^B. Similarly, the indifference condition asserts that there must be a convex combination of x^A and x^0 to which x^B is indifferent, and such a convex combination is Pareto inferior to x^A by definition of x^0; we denote it as $x^{B'}$. Since the substitution condition allows us to replace $x^{A'}$ and $x^{B'}$ with x^A and x^B, respectively, in any comparisons of outcomes, this establishes continuity of $\succsim_{1'}$.

Transitivity: Consider $x^A, x^B, x^C \in X$ such that $x^A \succsim_{1'} x^B$ and $x^B \succsim_{1'} x^C$. By completeness and transitivity of $\succsim_i$, there must be a weak ordering on $\{x_i^A, x_i^B, x_i^C\}$ for all i. Let x_i^* denote the most preferable of the three outcomes for individual i, and x_i^0 denote the least preferable. We now adapt x^* and x^0 slightly to incorporate three outcomes; they denote the vector of x_i^* and the vector of x_i^0, respectively. As previously, the indifference condition implies the existence of convex combinations $x^{A'}, x^{B'}, x^{C'}$ of x^* and x^0 such that $x^{A'} \sim_{1'} x^A, x^{B'} \sim_{1'} x^B$, and $x^{C'} \sim_{1'} x^C$. By the substitution condition, since $x^A \succsim_{1'} x^B$ and $x^B \succsim_{1'} x^C$, it must be true that $x^{A'} \succsim_{1'} x^{B'}$ and $x^{B'} \succsim_{1'} x^{C'}$, and thus by definition of x^* and x^0, $x^{A'}$ is Pareto superior to $x^{A'}$, and $x^{B'}$ is Pareto superior to $x^{C'}$. Therefore, $x^{A'}$ is Pareto superior to $x^{C'}$, which implies $x^{A'} \succsim_{1'} x^{C'}$. By the substitution condition, $x^A \succsim_{1'} x^C$, which establishes transitivity.

Proof of Theorem 2: Let $\succsim_{1'}, \ldots, \succsim_{n'}$ satisfy midvalue consistency. We can express the altruistic value functions as $V_i(x_1 \ldots, x_n) = \sum_{j=1}^{n} w_{ij} v_{ij}(x_j)$ for all i. Let x_{-k} denote a vector of individual outcomes for everyone except Person k. Arbitrarily, let (x_k^A, x_k^B) have a common tradeoff midvalue x_k^M for Person k. That means that for all individuals i, there exist $(x_{-k}^{iL}), (x_{-k}^{iH})$ such that $(x_{-k}^{iH}, x_k^A) \sim_{i'} (x_{-k}^{iL}, x_k^M)$ and $(x_{-k}^{iH}, x_k^M) \sim_{i'} (x_{-k}^{iL}, x_k^B)$. Expressing these indifference relations via the altruistic value functions yields, for Person i:

$$w_{ik}v_{ik}(x_k^M) + \sum_{j \neq k} w_{ij}v_{ij}(x_j^{iL}) = w_{ik}v_{ik}(x_k^A) + \sum_{j \neq k} w_{ij}v_{ij}(x_j^{iH}),$$
$$w_{ik}v_{ik}(x_k^B) + \sum_{j \neq k} w_{ij}v_{ij}(x_j^{iL}) = w_{ik}v_{ik}(x_k^M) + \sum_{j \neq k} w_{ij}v_{ij}(x_j^{iH}).$$

This pair of equalities can be reduced to:

$$v_{ik}(x_k^M) = 0.5v_{ik}(x_k^A) + 0.5v_{ik}(x_k^B).$$

Similarly to the $n = 2$ case, because x_k^A and x_k^B (and k) were chosen arbitrarily and the altruistic preference relations are continuous, infinitesimally similar levels will always have a tradeoff midvalue (i.e. these equalities can be established), and therefore the single-attribute value functions for any particular individual's outcome must all be positive linear transformations of one another.

We will now establish the converse. To ease the notational burden, we will consider the preferences of Person 1 and Person 2, without loss of generality. Let tradeoff midvalues x_k^M and $x_k^{M'}$ exist for Person 1 and Person 2, respectively, for arbitrary (x_k^A, x_k^B), and let $v_{2k} = cv_{1k}$, where $c > 0$. Then, for some $x_{-k}^{1L}, x_{-k}^{1H}, x_{-k}^{2L}$ and x_{-k}^{2H}:

$$w_{1k}v_{1k}(x_k^M)+\sum_{j\neq k}w_{1j}v_{1j}(x_j^{1L})=w_{1k}v_{1k}(x_k^A)+\sum_{j\neq k}w_{1j}v_{1j}(x_j^{1H})$$

$$w_{1k}v_{1k}(x_k^H)+\sum_{j\neq k}w_{1j}v_{1j}(x_j^{1L})=w_{1k}v_{1k}(x_k^M)+\sum_{j\neq k}w_{1j}v_{1j}(x_j^{1H})$$

$$w_{2k}cv_{1k}(x_k^M)+\sum_{j\neq k}w_{2j}v_{2j}(x_j^{2L})=w_{2k}cv_{1k}(x_k^A)+\sum_{j\neq k}w_{2j}v_{2j}(x_j^{2H})$$

$$w_{2k}cv_{1k}(x_k^H)+\sum_{j\neq k}w_{2j}v_{2j}(x_j^{2L})=w_{2k}cv_{1k}(x_k^M)+\sum_{j\neq k}w_{2j}v_{2j}(x_j^{2H})$$

After combining and canceling out common terms (note that c disappears as it did when $n=2$), we obtain:

$$v_{1k}(x_k^M)=0.5(v_{1k}(x_k^A)+v_{1k}(x_k^B))$$

$$v_{1k}(x_k^{M'})=0.5(v_{1k}(x_k^A)+v_{1k}(x_k^B)),$$

or $v_{1k}(x_k^M)=v_{1k}(x_k^{M'})$. By definition of v_{1k}, $x_k^M\sim_1 x_k^{M'}$, and the substitution condition ensures that $x_k^{M'}$ is also a tradeoff midvalue for Person 1. (If v_{1k} is invertible, then $x_k^{M'}=x_k^M$.) Since the choices of x_k^A and x_k^B (and k) were arbitrary, and the analogous argument can be made for any other individual, we can conclude that $\succsim_{1'},\ldots,\succsim_{n'}$ are midvalue consistent.

Derivation of (8): In this case, we have $2n$ variables $(V_1,\ldots,V_n,v_1,\ldots,v_n)$, and would like to express $V_1,\ldots,V_n$ as functions of $v_1,\ldots,v_n$. By moving all terms in the n general altruistic value functions to the left-hand side, we obtain:

$$V_i-f_i(v_i,V_1,\ldots,V_{i-1},V_{i+1},\ldots,V_n)=0$$

for $i=1,\ldots,n$. To apply the implicit function theorem, we require the matrix D of partial derivatives of these n equations with respect to $V_1,\ldots,V_n$. This matrix is given by:

$$D=\begin{bmatrix} 1 & -\frac{\partial f_1}{\partial V_2} & \cdots & -\frac{\partial f_1}{\partial V_{n-1}} & -\frac{\partial f_1}{\partial V_n} \\ -\frac{\partial f_2}{\partial V_1} & 1 & & & \\ \vdots & & \ddots & & \\ -\frac{\partial f_{n-1}}{\partial V_1} & & & 1 & -\frac{\partial f_{n-1}}{\partial V_n} \\ -\frac{\partial f_n}{\partial V_1} & & & -\frac{\partial f_n}{\partial V_{n-1}} & 1 \end{bmatrix}.$$

As in the $n=2$ case, the implicit function theorem states that if the determinant of D is non-zero, then a local functional representation of $V_1,\ldots,V_2$ in terms of $v_1,\ldots,v_n$ exists, as stated in (8).

Proof of Theorem 3: Let group homogeneity be satisfied. We know that preferences satisfy conditions that allow for the following form of altruistic value functions:

$$V_i(x)=\sum_{j=1}^{n} w_{ij}v_j(x_j)+w_{ig}v_g(x_g)$$
$$V_g(x)=\sum_{j=1}^{n} w_{gj}v_j(x_j)+w_{gg}v_g(x_g).$$

It is a straightforward consequence of group homogeneity that, for all x^A, x^B:

$$\sum_{j\neq i} w_{ij}v_j(x_j^A)+w_{ig}v_g(x_g^A)\geq \sum_{j\neq i} w_{ij}v_j(x_j^B)+w_{ig}v_g(x_g^B)$$

if and only if:

$$\sum_{j\neq i} w_{gj}v_j(x_j^A)+w_{gg}v_g(x_g^A)\geq \sum_{j\neq i} w_{gj}v_j(x_j^B)+w_{gg}v_g(x_g^B),$$

and therefore that:

$$(w_{i1},\ldots,w_{ii-1},w_{ii+1},\ldots,w_{in},w_{ig})=k_i(w_{g1},\ldots,w_{gi-1},w_{gi+1},\ldots,w_{gn},w_{gg})$$

for some $k_i>0$, since these weight vectors are unique up to positive linear transformations. This allows us to rewrite Person i's altruistic value function as:

$$V_i(x)=w_{ii}v_i(x_i)+k_i\left[\sum_{j\neq i} w_{gj}v_j(x_j)+w_{gg}v_g(x_g)\right],$$

or:

$$V_i(x)=(w_{ii}-k_iw_{gi})v_i(x_i)+k_iV_g.$$

If we let $w'_{ii}=w_{ii}-k_iw_{gi}$ and $w'_{ig}=k_i$, then $V_i(x)$ has the form stated in Theorem 3. Note that the case where $w_{ii}-k_iw_{gi}=0$ arises iff $\succsim_{g'}$ and $\succsim_{i'}$ are equivalent. When this happens, Person i's outcome can effectively be "combined" with the group outcome, as no other individuals distinguish between the two. From this point on, we assume that no individual's altruistic preference relation is equivalent to $\succsim_{g'}$; that is, $w'_{ii}\neq 0$ for all i.

To obtain the desired form for $V_g(x)$, we first rewrite this equation as:

$$V_g(x) = \frac{1}{k_i} V_i(x) - \frac{1}{k_i}(w_{ii} - k_i w_{gi}) v_i(x_i).$$

Given that $V_i(x)$ can be expressed in the desired form as given above, to obtain the desired form for $V_g(x)$, it will suffice to show that there exist w'_{gg} and w'_{gj}, $j = 1, \ldots, n$ such that:

$$V_g = w'_{gg} v_g(x) + \sum_{j=1}^{n} w'_{gj} \left[(w_{jj} - k_j w_{gj}) v_j(x_j) + k_j V_g \right],$$

which can be rearranged to yield:

$$\left(1 - \sum_{j=1}^{n} w'_{gj} k_j \right) V_g = w'_{gg} v_g(x) + \sum_{j=1}^{n} w'_{gj} \left[(w_{jj} - k_j w_{gj}) v_j(x_j) \right],$$

or:

$$V_g = \frac{w'_{gg} v_g(x) + \sum_{j=1}^{n} w'_{gj} (w_{jj} - k_j w_{gj}) v_j(x_j)}{1 - \sum_{j=1}^{n} w'_{gj} k_j}.$$

It is straightforward to use the initial expression for V_g to show that this is indeed possible if a simple condition is met. Set w'_{gg} equal to w_{gg} and w'_{gj} equal to $w_{gj}/(w_{jj} - k_j w_{gj})$. The numerator is then equal to the initial expression for V_g; all that is required is:

$$\sum_{j=1}^{n} \frac{k_j w_{gj}}{w_{jj} - k_j w_{gj}} \neq 1.$$

This is precisely the condition imposed by the implicit function theorem. The summand is equal to:

$$\frac{\partial f_j}{\partial V_g} \frac{\partial f_g}{\partial V_j},$$

as given in (12), which does not assume an additive form. (Here, f is simply an additive function of the general altruistic values.) This summand can be crudely interpreted as a measure of the relative importance the group places on Person j's outcome as compared to the relative importance Person j places on her own outcome, where a positive number means Person j cares about her own outcome more than the group does. Analogously to the previous applications of the implicit function theorem, this condition states that the set of interactions between each individual and the group do not precisely cancel out.

The converse can be shown easily by substitution using the same approach as in previous proofs, establishing Theorem 3.

Derivation of (12): First, move all terms in (11) to the left side, yielding:

$$\begin{aligned} V_1 - f_1(v_1, V_g) &= 0 \\ &\vdots \\ V_n - f_n(v_n, V_g) &= 0 \\ V_g - f_g(v_g, V_1, \ldots, V_n) &= 0. \end{aligned}$$

Per the IFT, we need to construct the $n+1$ x $n+1$ matrix of partial derivatives. The partial derivative of the ith equation $(1 \le i \le n)$ with respect to V_i is 1. The partial derivative with respect to $V_j (i \neq j, 1 \le j \le n)$ is 0, since V_j does not appear in any individual value functions other than the jth one. The partial derivative with respect to V_g is $-\partial f_i / \partial V_g$. The partial derivative of the last equation with respect to V_i is $-\partial f_g / \partial V_i$, and the partial derivative with respect to V_g is 1. Therefore, the matrix of partial derivatives is:

$$\begin{bmatrix} 1 & 0 \ldots & 0 & -\frac{\partial f_1}{\partial V_g} \\ 0 & 1 & & \\ \vdots & \ddots & & \\ 0 & & 1 & -\frac{\partial f_n}{\partial V_g} \\ -\frac{\partial f_g}{\partial V_1} & & -\frac{\partial f_g}{\partial V_n} & 1 \end{bmatrix}.$$

That is, all entries are zero, except for the main diagonal, the rightmost column, and the bottom row. Again, we require that the determinant be non-zero. However, given this structure, we can simplify the expression for the determinant. The determinant of a k x k matrix D can be expressed as:

$$\sum_{\sigma} sgn(\sigma) \prod_{i} D_{i,\sigma(i)},$$

where the σ are all possible permutations of k elements, and $sgn(\sigma)$ is 1 for even permutations and -1 for odd permutations. (The identity permutation is considered even.) This is often overly complex to compute for large matrices. However, in the matrix above, nearly all of the permutations will result in $\prod_i D_{i,\sigma(i)} = 0$. There are only $n+1$ permutations for which this product is non-zero.

The identity permutation yields a product of 1. The other permutations are those that involve only one switch, where one of the elements is the $n+1$st element (these are odd permutations). The resulting determinant is:

$$1-\sum_{i=1}^{n}\frac{\partial f_i}{\partial V_g}\frac{\partial f_g}{\partial V_i},$$

so provided that

$$\sum_{i=1}^{n}\frac{\partial f_i}{\partial V_g}\frac{\partial f_g}{\partial V_i}\neq 1,$$

a local functional representation for $V_1,\dots,V_n,V_g$ exists.