Choice, Rationality and Welfare Measurement*

Jerry R. Green
Harvard University

Daniel A. Hojman
Harvard University

First Version: March 2007; This Version: October 2007

Abstract

We present a method for evaluating the welfare of a decision maker, based on observed choice data. Unlike the standard economic theory of revealed preference, our method is can be used whether or not the observed choices are rational. Paralleling the standard theory we present a model for choice such that the observations arise "as if" they were the result of a specific decision making process. However, in place of the usual preference relation whose maximization induces the observations, we explain choice as arising from a compromise among a set of simultaneously-held, conflicting preference relations. As in revealed preference theory, these simultaneously held preferences are inferred from the choice data and we use them as the basis to discuss the decision maker's welfare. In general our method does not yield a unique set of explanatory preferences and therefore we characterize all the explanatory sets of preferences. We use this set to compute bounds on welfare changes. We show that some standard results of rational choice theory can be extended to irrational decision makers. The theory can be used to explore a number of context-dependent choice patterns found in psychological experiments.
JEL Classification: D01, D11, D60
Keywords: welfare economics, quasi-rational, behavioral economics, psychology and economics, voting

*E-mails: jgreen@hbs.edu, Daniel_Hojman@ksg.harvard.edu

1 Introduction

The use of observed choice behavior to make inferences about welfare is one of the basic methods of economics. This classical "revealed preference" method is based on the assumption that choice is rational. Our objective is to extend this method to choice functions that are not rational, while following its fundamental logic and objectives as closely as possible.

The standard economic procedure consists of three steps. First there is given a set of choice problems, each of which is a subset of the set of all possible alternatives, and an observed choice made by a decision maker for each problem in this set. These problems and observations constitute the data. The data may or may not be complete, as some possible choice problems may not be included. It is assumed that the data contain no contradictions of rationality – no switches between observed choices when both remain available, and no possibility for indirectly inferring that choice is self-contradictory, such as the observation of cycles. The second step in the revealed preference methodology is where the key theorem lies. Provided that the set of choice problems is "rich enough", the theory tells us that a preference relation can be constructed with the property that the decision maker is behaving "as if" he or she were optimizing it. When the data are not sufficient to define a unique consistent preference relation, there is still a set of preferences relations that can explain them.[1] In all cases, revealed preference theory imagines that the data is generated as if some preference relation were being optimized. The third step uses this preference relation both to predict behavior "out of sample" and to measure welfare.[2] For choice problems outside the original set one can calculate the choice that would be made and one can evaluate whether or not this problem represents an improvement over any other choice problem. If the preference relation is given a cardinal representation, this numerical function can be used as a quantitative welfare measure and for the purpose of computing interpersonal compensations.

The assumption of rationality is, however, not valid for many if not most data sets that have been encountered. Tests of rationality on ordinary de-

[1] Afriat's procedure finds one of them (Afriat, 1961). Mas-Colell (1978) gives an approximation result: The more data the smaller the set of preferences that remain consistent with them.

[2] If there are multiple preference relations consistent with the data one should make predictions based on each of them, and measure welfare using each of them. This procedure is not followed in practice, however.

mand data fail not because of large error terms but because the hypothesis is demonstrably false.[3] Choice data in psychological experiments or in field-based observations also contain internal inconsistencies and contradictions. It is frequently the case that choice varies systematically with the context in which it is made, refuting any internal consistency axiom that might be applied.[4] Revealed preference theory cannot be used as a basis for welfare analysis because rationality cannot reasonably be assumed. There simply is no single preference relation that generates the data, and thus there is no preference relation that can serve as a basis for welfare measurement.

In this paper we retain the objective of constructing a model for the decision maker's observed choice. We seek a model that works whether or not these choice contain contradictions to rationality. Our strategy is to look for a set of preference relations and a method for aggregating them that work in the same "as if" sense that is employed in the standard theory. Our interpretation of this set of preference relations is that they represent multiple conflicting motivations that influence the decision maker's choice. Thus we model decision makers who are "conflicted" in that they simultaneously hold multiple preferences over the alternatives. We want the data to tell us what conflicts the decision maker might be experiencing; and we respect all these conflicting preferences when evaluating welfare. It is in this sense that we retain the central principle of choice-based welfare economics that has been the hallmark of microeconomics.

Our central definition is that of an "explanation": A set of preferences and an aggregation method such that this method, applied to these preferences, reproduces the observed choice at every datum.

Aggregating conflicting preferences is the domain of social choice theory, the most important result of which is Arrow's Theorem. This theorem tells us that any non-trivial aggregation of preferences must display inconsistencies. Democracy and collective rationality are incompatible. This is usually taken as a negative, disappointing, result. For our theory, however, this re-

[3] See, for example, Deaton-Muellbauer (1980). Rationality in demand is tested via the implication that under rational choice the Slutzsky substitution terms are symmetric.

[4] Rubinstein-Salant (2007) describe a theory of "choice with frames" which can explain some forms of inconsistency of this type. This approach is related to that in our paper in that they seek a more general theory that can incorporate irrationality and derive the nature of the irrationality from the data. On the other hand, in their theory there is one complete ordering that describes any particular choice datum, whereas in our theory the choice will turn out to be a compromise among the relevant orderings.

sult is a source of strength. Because we explain irrationality as a necessary consequence of compromise among conflicting objectives, we seek a type of converse to social choice theory: We ask what forms of conflict and compromise can play the same "as if" role in an explanation of an irrational choice pattern that a single preference relation plays for rational data in revealed preference theory?

Our results are as follows:

First we show that there are generally many explanations for any finite data set, and that they form a convex subset in the relevant space. When evaluating welfare using our method, we obtain a range of welfare measures due to the multiplicity of the explanations.[5] The geometric structure of the set of all explanations makes it feasible to compute bounds on welfare.

Second we show that any choice function, no matter how irrational, can be explained by our method. One may construe this as a negative result in that the theory is therefore not testable. We do not, however, view it in this light. Although there are no logically-imposed limits on the extent of the irrationality that can be encompassed by this theory, highly irrational choice functions may require explanations by very strange sets of simultaneously held preferences. If it were possible to rule out those combinations of preferences either by assumption or on the basis of other evidence, one could provide testable restrictions that even irrational choice must satisfy.

Third we address the central question of whether an expansion in the set of possibilities is necessarily better (or at least not worse) for the decision maker, as it would be for a rational person with a single well-defined preference relation. We show that although the decision maker may be irrational in some respects, it is possible to preserve the conclusions of rational choice theory provided that the irrationality does not affect the two alternatives that are chosen before and after the expansion. Specifically, suppose that the expansion causes a change in outcome from y to x and that y is never chosen at any situation where both are available. Even though the choice function may be highly irrational in other respects, the consistency of choice at it pertains to x and y guarantees that there exist explanations that give all the weight possible to preference in which x is indeed preferred to y. Thus the beneficial nature of expansions of the available set can be insulated from observed irrationalities that are not relevant to the expansion at hand.

[5] As mentioned above, revealed preference theory would display the same sort of multiplicity on finite data sets.

Fourth, we explore a converse to this result. Whenever there is any irrationality in behavior we can find a pair choice problems where choice switches from x to y when y is added to the available set, and yet there is some explanation of this behavior such that every one of the preferences that forms part of this explanation is hurt by this expansion.

Fifth we go beyond the above results which rely on only choice-based information. When some cardinal, non-choice based information is available, we can derive more detailed welfare conclusions. The idea is, again, to parallel rationality-based theory, which constructs cardinal measures of welfare based on assumptions beyond the rationality of the choices observed.[6] We show that the most optimistic and most pessimistic bounds on welfare are always realized within a special highly structured family of explanations that we characterize. Because of the complexity of the general set of explanations for choice functions, this result drastically simplifies the calculation of the best and worst case scenarios for welfare change. As an application of this theory we show what it implies in the case of three alternatives – a case that has been thoroughly documented in psychological experiments.

The accumulation of laboratory and field-based evidence of irrationality over the past twenty years has led other authors to explain choice either by sequential procedures involving multiple objectives, or as the outcome of a game in which these conflicting objectives play strategically against each other. Papers in the non-strategic, sequential category posit additional information about the alternatives in order to structure the procedure. For example, the alternatives may be described by a list of attributes which could be considered in a fixed order to eliminate or reorder the alternatives. Observed choice is explained "as if" this process were followed by the decision maker. From a welfare point of view, however, it is not clear which of the attributes is most salient – or, if preferences over attributes are to be combined, how should they be weighted. Classic studies in this mode are Tversky

[6] These assumptions are of several types. One uses studies of brain function, other physiometric measures, or self-reported measures of satisfaction as the basis for the cardinalization. Another uses further data on choices among lotteries, and then, under the assumption that these choices fulfill an independence condition, produces a cardinal utility from the observed risk preferences. Finally, if the choices are over commodity bundles and one of the commodities, typically money, is assumed to enter preferences in a quasi-linear form, then this commodity can be used to perform the scaling – yielding the "money metric" utility scale. All of these methods thus rely on additional axioms and assumptions to produce acardinalization.

(1972), Shafir (1993), and Shafir-Simonson-Tversky (1993). More recent papers along the same lines are Ahn-Ergin (2007) and Manzini-Mariotti (2007).

Models of irrational choice in the game theoretic category recognize the existence of multiple conflicting preferences and impose strategic structure on the problem within which these preferences interact. These studies are attractive because many people do consciously question their own motivations, engage in introspection, and sometimes consciously act so as to suppress motivations that they deem harmful to their own true interests. The difference between these models and our is that they take the nature of the multiple selves to be exogenous, using a fixed structure of these selves to generate the irrational observations. We, on the other hand, ask the data to generate the multiple explanatory preferences for us. When the nature of the decision problem has enough structure that one can identify classical motivations, such as patience and impulsiveness, the strategy followed by these papers can pay handsome dividends. Our structure is perhaps more appropriate in general context-dependent situations where the detailed nature of the alternatives is unknown to the analyst and a priori assumptions about which preference plays which strategic role in the game may not be appropriate. Important papers in this strategic mold include Strotz (1956), Schelling (1984), Bernheim-Rangel (2004), Gul-Pessendorfer (2001), and Fudenberg-Levine (2006). In many of these papers the pattern of choice that is generated by game theoretic interactions among the preferences displays enough consistency to be "explained" as the maximization of a composite, aggregated preference. These papers do not address the question of how to measure welfare. It would seem that one either would have to accept the aggregated preference as the appropriate welfare measure, or as in the case of the sequential procedures, welfare analysis would depend upon making one of the underlying preferences more salient than the others.[7]

Our paper is, to our knowledge, the first to take a non-strategic approach to the problem of aggregating conflicting motivations. We treat both the alternatives and the motivations symmetrically, imposing no structure on the characteristics of the alternatives.. We take this symmetric approach in order to allow the observed choice behavior to be the sole determinant of the welfare analysis.

[7] Sometimes these papers adopt a specific choice for the aggregation of conflicting preferences, such as Laibson, Repetto and Tobacman (1998) which gives priority to the long-run self rather than any of the more impatient selves. Se also O'Donahue and Rabin(1999).

The paper is organized as follows: The basic set up, notation and concepts are given in section 2. Section 3 is devoted to a three-alternative example, typical of psychological evidence on the presence of context effects, involving choice among delayed payoffs. Section 4 gives an analysis of the general three alternative case. In Section 5 we deal with the general case, showing that the results for three alternatives generalize without much change. Section 6 is devoted to the question of how a limited, weak form of rationality concerning the comparison of two particular alternatives, which we call pairwise coherence, allows some fundamental welfare conclusions to survive in the presence of considerable irrationality of the choice function more generally. Through Section 6 we are making only ordinal comparisons. In section 7 we allow for each explanatory preference to be given a cardinal representation, and we adopt a utilitarian criterion when aggregating multiple conflicting preferences in an explanation. In this framework we evaluate welfare bounds for expansions of the sets of alternatives and for some particular well-documented irrational behavior patterns in the three alternative case. A brief concluding section follows. Proofs are in the Appendix.

2 Choice and Welfare Measurement

2.1 Conflicting Motivations and a Voting Model of Choice

The set of all possible *outcomes* is X. A typical outcome is $a \in X$. A set of available alternatives $A \subseteq X$ is a *choice situation*. An *observation* is a pair (a, A) with $a \in A$. We observe choices from a domain $\mathcal{A}$ of choice situations A. Thus the data we need to explain is a *choice function* $c : \mathcal{A} \to X$ summarizing all the observations $(c(A), A)$ for $A \in \mathcal{A}$. We will take $\mathcal{A}$ to be the set of all non-empty subsets of A unless otherwise noted.[8]

The set of all strict orders on X is denoted Π. An individual is identified with a probability distribution $\lambda \in \Delta^{\Pi}$, the set of all probability measures over Π. Our interpretation of λ is that it describes the simultaneously held

[8] One of the strengths of revealed preference theory is that it uses the structure of the available sets, in particular the linear structure of consumers' budget sets, to make indirect inferences about preferences. This enables the construction of a preference relation from families $\mathcal{A}$ that are far smaller than the set of all non-empty subsets of X. Moreover, Forges and Minelli [2006], and Fostel, Scarf and Todd [2004] have extended standard revealed preference theory to the case of non-linear budget sets, and have shown how sparse $\mathcal{A}$ can be while still enabling the construction of an explanatory preference relation.

preferences that motivate an individual, as well as the strengths of each of these motivations. For this reason we use the term *motivations* when speaking about the orders π. We call λ a *population of motivations* or, for brevity, a *population*. An individual's choice behavior is to be explained as if the population λ were aggregated by some fixed procedure. Let $v : \Delta^{\Pi} \times \mathcal{A} \to X$ describe this procedure. We will call v a *voting rule*. The interpretation of a voting rule is that if the population of motivations is λ and the available set of alternatives is A then $v(\lambda, A)$ is the result of the vote by this population over the set A. Further specifications and restrictions on the voting rules that we consider will be discussed below in section 2.2.

Thus, we are given a choice function c and seek to explain c "as if" it were generated by an individual who is characterized by a pair (λ, v).

Definition 1 *An* explanation *of a choice function c is a pair (λ, v) consisting of a population λ and a voting rule v such that $c(A) = v(\lambda, A)$ for all $A \in \mathcal{A}$.*

For a fixed voting rule v the set of populations λ such that (λ, v) is an explanation of c is denoted $E(c, v)$. If the voting rule were not restricted in some way any choice function can be "explained", and any population λ can be part of that explanation. One could simply let the voting rule ignore λ and chose $c(A)$ whenever the available set is A. Thus all interesting conclusions of our model are driven by the restrictions that we place of the form of the voting rule v.

Let us restrict the voting rule by requiring that v lie within a specified family of rules V. The smaller V is, the fewer explanations of c there will be. Thus it becomes interesting to ask, for a particular family V, whether a given choice rule c can be explained by any (λ, v) with $v \in V$. If such explanations do exist, the populations λ that are part of these explanations will reflect the restriction to $v \in V$ as well as the rationality or irrationality of c. For example, if c actually satisfies the axiom of revealed preference, and V includes voting rules that respect unanimity ($c(A)$ is the maximal element of π on A whenever λ is a point mass at π), then there will be a "rational explanation" of c. There may, however, be other explanations of c as well. To take one more example, if $v \in V$ were Plurality rule ($c(A)$ is the element that maximizes the weight on the first choice of π within A, under the distribution λ), then a cyclic choice pattern c will be explainable by (λ, v) provided that λ admits the required Condorcet paradox.

In this paper we explore explanations that are based on a particular family of voting rules, the scoring rules V_s, such as Borda count or plurality. Further details on the construction of scoring rules are given shortly in Section 2.2.[9]

Definition 2 *A* scoring rule explanation *of c, is an explanation of c, (λ, v), where $v \in V_s$.*

Given a choice function c, the set of populations that are associated with some scoring rule explanation of c is $E(c) = \cup_{v \in V_s} E(c, v)$. This is the set $\lambda \in \Delta^{\Pi}$ such that (λ, v) explains c over $\mathcal{A}$ for some $v \in V_s$. For the family of scoring rule explanations of this paper we are interested primarily in three questions. Which choice functions c have explanations? And for a given choice function c, what is the set $E(c)$? We provide answers to these questions in sections 4 and 5. Finally, given that we have restricted the explanatory populations to lie in $E(c)$, what can be said about the change in the decision maker's welfare when the set of available alternatives varies? We approach this welfare question in sections 6 and 7.

Decision rules that derive from voting procedures, particularly those that derive from scoring rules, are interesting for a number of reasons. Scoring rules ignore cardinal information: only ordinal information is necessary in order to determine voting outcomes. Thus any difference between the welfare maximizing outcome and the actual voting outcome can be traced to the fact that voting via scoring rules ignores the intensities of motivations that a true welfare maximization would require.[10]

2.2 Scoring Rules and Explanations

A *scoring rule* $v \in V_s$ is characterized by a set of $|X| - 1$ scoring vectors $\{\gamma_k\}_{k \in \{2,...,|X|\}}$, where γ_k is the scoring vector that applies when the available set has k alternatives. We write $v = (\gamma_2, \gamma_3, ..., \gamma_{|X|})$ and the k-alternative scoring vector

$$\gamma_k = (\gamma_k^1, \gamma_k^2, ..., \gamma_k^k)$$

[9] It is also possible to specify V_s axiomatically, as in Young (1975).

[10] However we note that purely ordinal theory forms one endpoint of the continuum of preference aggregation procedures. In a second paper we study the relationships between the performance of this ordinal theory and more powerful methods for aggregation that use increasingly some of the cardinal intensity information to compare the influence of different motivations. See Green and Hojman [2007].

has k components satisfying $\gamma_k^1 \geq \gamma_k^2 \geq ... \geq \gamma_k^k$, and at least one these inequalities is strict. The scoring vector gives the number of "points" γ_k^j assigned to the j^{th} ranked alternative among the k alternatives in a subset A. Without loss of generality we assume $\gamma_k^1 = 1$ and $\gamma_k^k = 0$ for all $k \in \{2, ..., |X|\}$. Given a choice situation $A \subseteq X$, $k = |A|$, and order π that ranks alternative a as the j^{th} best alternative among those available from A, the *score* of alternative a from A under the ordering π is $s(a, \pi, A) = \gamma_k^j$. The *total score* of alternative x in choice situation A given a population λ is then

$$s(a, \lambda, A) = \sum_{\pi \in \Pi} s(x, \pi, A)\lambda(\pi).$$

The result of a vote under v by a population λ when the set of alternatives is A is the alternative that receives the highest score, that is,

$$v(\lambda, A) = \arg\max_{x \in A} \; s(a, \lambda, A) = \sum_{\pi \in \Pi} s(a, \pi, A)\lambda(\pi).$$

Examples of scoring rules include Plurality, the Borda count, and Antiplurality. Scoring rules are a special case of general positional voting rules.[11]

Example 1 (Plurality) *Plurality is characterized by weights $\gamma_k^j = 0$ for all $k \in \{2, ..., |X|\}$, $j \in \{2, ..., k\}$. Thus, scoring vectors have the form $s_k = (1, 0, ..., 0)$.*

Example 2 (Borda) *The Borda count is characterized by weights $\gamma_k^j = \frac{k-j}{k-1}$, $k \in \{2, ..., |X|\}$, $j \in \{1, ..., k\}$.*

Example 3 (Antiplurality) *Antiplurality is characterized by weights $\gamma_k^j = 1$ for all $k \in \{2, ..., |X|\}$ and $1 \leq j < k$, and $\gamma_k^k = 0$.*

For any scoring voting rule $\gamma_2 = (1, 0)$, so that when deciding between only two available alternatives, all scoring rules are identical. Note also that, in the case of three alternatives, a scoring rule v is determined by a single parameter, the weight γ_3^2 assigned to the second best choice from three alternatives.

[11] General positional voting rules are those rules that depend on the rank orders of the alternatives. In scoring rules this dependence is restricted to be additive – resulting in the scores $s(a, \lambda, A)$.

We conclude this section with an important remark. For a fixed scoring rule and a choice function c we can rewrite

$$c(A) = v(\lambda, A) = \arg\max_{x} \sum_{\pi \in \Pi} s(a, \pi, A)\lambda(\pi)$$

as a set of choice inequalities

$$\sum_{\pi \in \Pi} s(c(A), \pi, A)\lambda(\pi) \geq \sum_{\pi \in \Pi} s(a, \pi, A)\lambda(\pi) \text{ for all } a \in A. \tag{1}$$

A key property of these inequalities is their *linearity* in the population strengths vector λ. Each inequality defines a half space in the space of populations Δ^{Π}. As a result, the set of explanations $E(c, v)$ for a fixed voting rule v is the *polytope* defined by the set of these linear inequalities for each $a \in A$ and each $A \in \mathcal{A}$.[12] Thus, the set $E(c) = \cup_{v \in V_s} E(c, v)$ of all populations that can be part of an explanation is a union of polytopes, which is not necessarily itself a polytope. We explore the geometric structure of $E(c)$ and $E(c, v)$ for $v \in V_s$ extensively below.

2.3 Welfare Measurement

We use the set of explanations to derive conclusions about welfare changes as the available set varies. Initially, in sections 4, 5, and 6, we restrict ourselves to ordinally-based conclusions, by which we mean that all the explanatory motivations must agree on the welfare evaluation of the outcome. Any motivations that do not share this evaluation appear in the explanation of the choice function with a zero weight. Later in the paper, in section 7, we examine the case of cardinal preferences, which as mentioned above will need to be based on non-choice based information. When such information is available we choose to measure welfare based on the natural utilitarian criterion – adding the welfare changes across the population of preferences and weighting according to the relevant λ. In this case we will say that welfare is improved if the utilitarian sum increases for any explanation in $E(c)$.[13]

Formally speaking, we approach welfare measurement in the following sequence of steps:

[12] A polytope is a bounded polyhedron. A polyhedron is an intersection of half spaces. Since the simplex is bounded, so is the polyhedron defined by choice inequalities (1).

[13] Roemer (1996) has offered a powerful argument against the general applicability of the utilitarian criterion in interpersonal contexts.

(W1) Given a domain $\mathcal{A}$ of choice instances, observe $\{c(A)\}_{A\in\mathcal{A}}$.

(W2) Given a family of voting rules V compute the set of populations $E(c,V)=\cup_{v\in V}E(c,v)$ that can be part of an explanation.

When only ordinal, choice-based, information is to be used in the welfare analysis, we then proceed to step (W3, Ordinal)

(W3) (*Ordinal*) Given a change in the set of opportunities from A to B, test whether there exists $\lambda\in E(c,V)$ such that $c(A)$ and $c(B)$ are ranked in the same way by all π for which $\lambda(\pi)>0$.

For a given change in the available set, step (W3) focusses our attention on the subset of choice functions for which unambiguous welfare evaluations are possible, defined as follows.

Definition 3 *Let $F=(A,B)\in\mathcal{A}\times\mathcal{A}$ denote a change in the opportunity set from A to B. A choice function c is said to* admit a positive welfare inference at F *if there exists $\lambda\in E(c,V)$ such that there is unanimity with regard to the ranking of $c(A)$ and $c(B)$ among the preferences π that are given strictly positive weight $\lambda(\pi)>0$.*

The set of all choice functions admitting a positive welfare inference at F is denoted by $\Delta W^{+}(F)$.

Alternatively, if cardinal preference information is available as a basis for welfare measurement it can be summarized by $u:X\times\Pi\to\mathbb{R}$, which defines a cardinal utility function $u(\cdot,\pi)$ for each $\pi\in\Pi$. Then it is natural to compute utilitarian welfare functional $\widetilde{W}$ that aggregates the cardinal information.

Definition 4 *The utilitarian welfare functional $\widetilde{W}$ based on the cardinalization u is*

$$\widetilde{W}(x,\lambda,u)=\sum_{\pi\in\Pi}u(x,\pi)\lambda(\pi) \tag{2}$$

If cardinal preference information is available we do not need the unanimity criterion of step (W3). Welfare can be evaluated directly:

(W3') (*Cardinal*) For each $\lambda\in E(c,V)$ compute the welfare measure at A based on λ:

$$W(A,c,\lambda,u)=\widetilde{W}(c(A),\lambda,u). \tag{3}$$

Of course, as there are multiple explanatory populations the sign of welfare change may depend on this population. For any λ the change in welfare associated to the change in opportunities $F = (A, B)$ is

$$\Delta W(F, c, \lambda, u) = W(B, c, \lambda, u) - W(A, c, \lambda, u). \tag{4}$$

3 A Motivating Example

To make our ideas concrete, we consider a motivating example that concerns the choice over consumption that is received after a known time delay. Each alternative a is a particular consumption level m_a and an associated delay t_a. We will write $a = (m_a, t_a)$. We use this example to illustrate the welfare measurement method proposed above in a familiar economic context.

Take three alternatives $x = (10, 0)$, $y = (15, 1)$ and $z = (35, 2)$. The choice function defined on subsets of these three alternatives will be cyclical on pairwise choices, and will select x when all three are available:

$$c(\{x, y\}) = x,\ c(\{y, z\}) = y,\ c(\{x, z\}) = z,\ \text{and } c(\{x, y, z\}) = x.$$

This pattern is typical of experimental evidence and is a simplification of examples in Tversky (1969) and Roelfsma and Reed (2000)

To shorten the notation slightly we write xyz to indicate that the preference ordering π under consideration ranks x above y and y above z. Thus the six possible different orderings π_i, $i = 1, ..., 6$ are:

$$\begin{aligned} \pi_1 &= xyz, \quad \pi_2 = xzy \\ \pi_3 &= yzx, \quad \pi_4 = yxz, \\ \pi_5 &= zxy, \quad \pi_6 = zyx \end{aligned}$$

One can generate all of these preferences from a standard discounted utility formulation $u_i((m, t)) = \tilde{u}_i(m)e^{-\varrho t}$ in which the Bernoulli utility $\tilde{u}_i(m)$ function varies but the time preference parameter ϱ is common across the six motivations. Moreover the Bernoulli utilities can be chosen so that they are concave in m.[14]

Now let us examine some of the explanations that exist for the cyclical c that has been observed. Given the population, when the set of alternatives

[14] See Appendix A.

has only two elements, all voting rules produce the same result . Thus, to produce a cycle on the three two-element sets, the population must be display a Cordorcet pattern. For example, $\lambda = (\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0)$ produces pairwise votes of $\frac{2}{3}$ to $\frac{1}{3}$ with the majority in favor of the indicated choice in each instance. Other populations produce the same cyclic choice patterns, but with different majorities of the motivations in favor of the winner. A borderline case is the population $\lambda = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0)$ where x is unanimously preferred to y and there is a tie on the other two two-element sets. Now we consider the one additional piece of evidence – the fact that x is chosen from the three-element set. To be part of an explanation, the population must also lead to this choice, and here the voting rule is relevant to the outcome. If we examine Plurality rule, for example, then $\lambda = (\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0)$ produces a three-way tie on this set, and $\lambda = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0)$ produces a tie between x and z. Such ties can be avoided at voting rules other than Plurality. At Antiplurality or Borda, $\lambda = (\frac{1}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}, 0)$ still produces a tie, by symmetry, but $\lambda = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0)$ produces x as the unique outcome because z is last in π_1 but x is second in π_5.

The case of $\lambda = (\frac{1}{2}, 0, 0, 0, \frac{1}{2}, 0)$ is quite interesting because arbitrarily close to this population are the populations $\lambda = (\frac{1-\epsilon}{2}, 0, \epsilon, 0, \frac{1-\epsilon}{2}, 0)$. These populations, together with the appropriate non-Plurality voting rules, uniquely select all the required choices. Moreover, as ϵ can be made arbitrarily small, almost all the weight is placed on preferences that favor x over y.

Let us now consider the evidence presented by this choice function as to the pairwise preference that might be held by this decision maker. For the pair $\{x, y\}$ the evidence is unambiguous that $x > y$. Whenever both x and y are available, x is selected – independent of the presence or absence of z. For the pair $\{x, z\}$ there is contradictory evidence since the presence of y causes x to be chosen, but z is chosen in y's absence. For the pair $\{y, z\}$ the evidence favors $y > z$, but as we do not know what the second choice would be from $\{x, y, z\}$ the evidence for this preference is weaker than that for x versus y.

These pairwise considerations are relevant when we consider the expansion in the set of alternatives. We would like to be able to draw inferences as to whether an expansion is the set of alternatives is good for the decision maker, as it would be for someone who is rational. Consider the expansion from $A = \{y, z\}$ to $X = \{x, y, z\}$ and the consequent change in the choice from $y = c(A)$ to $x = c(X)$. Is there an explanation consistent with the evidence that x is always chosen in preference to y? We can answer this

question in the affirmative by considering the explanation of c via the population $\lambda = (\frac{1-\epsilon}{2}, 0, \epsilon, 0, \frac{1-\epsilon}{2}, 0)$. In this explanation there is near unanimity that this expansion is beneficial. Thus we see that the irrationality of cyclic choice does not stand in the way of some explanations which very strongly indicate that the switch from y to x is beneficial. This example suggests, therefore, that a revealed preference between two alternatives can be consistent with as large a majority in favor of the better one as one would like to have, despite any revealed irrationality with respect to alternatives other than these two. We will see in Theorem 3 that this is a general proposition.

The other two expansions lead to different conclusions. Going from $\{x, y\}$ to $\{x, y, z\}$ is irrelevant to welfare, as x is chosen in both instances. Going from $\{x, z\}$ to $\{x, y, z\}$ switches the choice from z to x. One can show, however, that there is no explanation of c that put all the weight, or approximately all the weight, either on preferences concentrated on $x > z$ or on preferences concentrated on $z > x$. Non-trivial conflict is a necessary ingredient in any explanation of this choice function, as far as the pair $\{x, z\}$ is concerned. Therefore, as we are restricting ourselves to ordinal information about the motivations, there is no explanation from which we can decide which of these alternatives is preferred. The fact that this type of ambiguous welfare conclusion is an unavoidable consequence of irrational choice patterns is a general proposition – see Theorem 4.

Now let us turn to the case in which cardinal welfare measures are available for each of the motivations. Cardinal information in our model must come from observations beyond the choice function itself. We need such cardinal information if we are going to make welfare comparisons in ambiguous cases such as that of x versus z above. For concreteness, let us use as this cardinal measurement the parameters of Appendix A in the Bernoulli utility that represents each motivation.

For these parameters we can see that quantitative welfare measurement is ambiguous in sign even for an expansion for which there is unambiguous evidence of pairwise preference – as in the comparison of x to y. This can be shown by considering the following two explanations of c: (λ, v^P), with $\lambda = (.43, 0, .40, 0, .17, 0)$ and v^P is Plurality rule, and (λ', v^A), with $\lambda' = (.40, 0, .17, 0, .40, .03)$ and v^A is Antiplurality. $W(\{y, z\}, c, \lambda, u) = 1.176$ and $W(\{x, y, z\}, c, \lambda, u) = 1.165$, yielding a decrease in utilitarian welfare On the other hand, for the population λ' we obtain $W(\{y, z, \}, c, \lambda', u) = 1.149$ and $W(\{x, y, z\}, c, \lambda', u) = 1.174$, so that the welfare change for the explanation λ' is positive. Thus in this example the sign of the cardinal welfare

evaluation for the expansion from $\{y, z, \}$ to $\{x, y, z\}$ is ambiguous, even though the decision maker has revealed no irrationality with respect to the two alternatives that are chosen before and after the expansion. We return to this issue in sections 6 and 7.

4 Choice Rules in the Three Alternative Case

We provide a classification of choice rules and a characterization of the choice functions that can be explained by our theory above for the case in which X has three elements. The results for the general case are presented in Section 5. For concreteness, let $X = \{x, y, z\}$.

4.1 The Four Choice Rules

If X has three alternatives there are four possible choice situations, three corresponding to choices from two-element subsets of X and one for choice over the entire set. This yields twenty-four possible choice rules for the three-alternative case[15] which can be classified into four choice patterns: *Seemingly Rational Choice*, *Second Place Choice*, *Third Place Choice* and *Cyclic Choice*. Rules in the same class are formally identical modulo a permutation of the alternatives.

The first three classes of choice rules have in common the existence of an alternative x^r which is "Condorcet winner" in pairwise contests. Let x^s and x^t denote the other alternatives, where $x^s = c(X \backslash \{x^r\})$ is the "second-place" choice and x^t is the "third-place" choice, $x^t \notin \{\{x^r, x^s\}$. When pairwise choice exhibits a Condorcet winner in this three-alternative environment, pairwise choice will have "revealed" a preference relation $x^r \succ x^s \succ x^t$. For example, if $X = \{x, y, z\}$ the following choice function

$$c(\{x, y\}) = x,\ c(\{x, z\}) = x,\ c(\{y, z\}) = y$$

has $x^r = x$, $x^s = y$ and $x^t = z$ which is consistent with the order "x over y over z". The question remaining, however, is whether this preference relation also characterizes the choice when all three alternatives are available.

The three classes of choice rules are defined by the choice from the triple. If $c(X) = x^r$ the choice rule is *Seemingly Rational*. If $c(X) = x^s$ the choice

[15] There are three two-alternative choice instances and the choice instance corresponding to the triple give $24 = 2 \times 2 \times 2 \times 3$.

rule displays *Second Place Choice*. If $c(X) = x^t$ the choice rule displays *Third Place Choice*. Clearly, second and third place choice are not consistent with explaining choice based on the maximization of a single preference relation.

When pairwise choice does not exhibit a Condorcet winner, *Cyclic Choice,* the fourth class of choice rules, exists. For example,

$$c(\{x, y\}) = x, \ c(\{x, z\}) = z, \ c(\{y, z\}) = y. \tag{5}$$

This example exhibits the cycle "x over y over z over x".[16]

We conclude this section remarking that the three patterns inconsistent with rational behavior have been documented by the experimental psychology and decision-making literature that focuses on the context effects that arise in choice with multi-attribute alternatives. A classic paper by Tversky (1969) and more recent work by Roelofsma and Read (2000) show that cyclic choice can arise systematically. There is also robust evidence of Second Place Choice, as shown by Simonson (1989). Third place choice seems to be more elusive but Redelmeier and Shafir (1995) finds this pattern.[17]

4.2 Characterizing Explanations in the Three-Alternative Case

In the case of three alternatives the set of all orderings Π has six elements. As in Section 3, we adopt the concise notation that, for a generic $\pi \in \Pi$, if a_1 is preferred to a_2 and a_2 is preferred to a_3 we write $\pi = a_1a_2a_3$. With this notation the six strict orderings are

$$\begin{aligned} \pi_1 &= xyz, \quad \pi_2 = xzy \\ \pi_3 &= yzx, \quad \pi_4 = yxz, \\ \pi_5 &= zxy, \quad \pi_6 = zyx \end{aligned}$$

[16]The absence of a Condorcet winner necessitates the existence of a pairwise cycle with three or more elements. Consider the directed graph in which each vertex is an alternative in X and there exists a directed edge from a vertex a to another vertex a' if $c(\{a, a'\}) = a$. If c does not have a Condorcet winner this means that each vertex has an outgoing edge and, since there are a finite number of edges, the graph must contain a directed cycle.

[17]The prevailing psychology theories include sequential decision-making procedures such as *elimination by aspects* or theories based on context-dependent salience such as *asymmetric dominance*. A more comprehensive theory called *reason-based choice* is proposed by Shafir, Simonson, and Tversky (1993). This theory, based on the idea that the context determines which among of many conflicting reasons prevails in a given choice situation, is close in spirit to the model presented in this paper.

and a population of motivations $\lambda \in \Delta^6$ is a probability distribution over these six orderings. For short, we write $\lambda_i = \lambda(\pi_i)$ for each $i = 1, 2, ..., 6$.

As mentioned earlier, with three alternatives a positional voting procedure v is determined by a single free parameter $\gamma_3^2 \in [0, 1]$ and therefore, we can identify the voting rule with this parameter: $v = \gamma_3^2 \in [0, 1]$. We are interested in characterizing the set of choice functions that can be explained by these rules. As illustrated by the following examples, given a positional voting rule v, if (λ, v) is an explanation of a choice function c then five choice inequalities must be satisfied.

Example 4 *Explaining Seemingly Rational Choice*

Consider the following Seemingly Rational choice function

$$c(\{x, y\}) = x,\ c(\{x, z\}) = x,\ c(\{y, z\}) = y,\ c(\{x, y, z\}) = x. \tag{6}$$

For an explanation based on voting rule $v \in [0, 1]$ the choice inequalities in (1) translate into

$$\begin{aligned}
\lambda_1 + \lambda_2 + \lambda_5 &\geq \lambda_3 + \lambda_4 + \lambda_6 && \text{(7a)}\\
\lambda_1 + \lambda_2 + \lambda_4 &\geq \lambda_3 + \lambda_5 + \lambda_6 && \text{(7b)}\\
\lambda_1 + \lambda_2 + \lambda_4 &\geq \lambda_3 + \lambda_5 + \lambda_6 && \text{(7c)}\\
\lambda_1 + \lambda_2 + v(\lambda_4 + \lambda_5) &\geq \lambda_3 + \lambda_4 + v(\lambda_1 + \lambda_6) && \text{(7d)}\\
\lambda_1 + \lambda_2 + v(\lambda_4 + \lambda_5) &\geq \lambda_5 + \lambda_6 + v(\lambda_2 + \lambda_3) && \text{(7e)}
\end{aligned}$$

Inequalities (6a)-(6c) correspond to choices from pairs of alternatives and express respectively that x beats y in the pairwise contest, x beats z in the pairwise contest, and y beats z in the pairwise contest. Inequalities (6d)-(6e) correspond to choice from the triple and express that x has a higher score than y and z when all three alternatives are available. $E(c, v)$, where c is the seemingly rational choice function, is the set of solutions to these five inequalities under the restriction that λ is in the five-dimensional simplex Δ^6.

Example 5 *Explaining Second-Place Choice*

Consider the Second-Place Choice rule

$$c(\{x, y\}) = x,\ c(\{x, z\}) = x,\ c(\{y, z\}) = y,\ c(\{x, y, z\}) = y.$$

An explanation of c based on v must satisfy inequalities (4a)-(4c) and

$$\begin{aligned}\lambda_3 + \lambda_4 + v(\lambda_1 + \lambda_6) &\geq \lambda_1 + \lambda_2 + v(\lambda_4 + \lambda_5)\\ \lambda_3 + \lambda_4 + v(\lambda_1 + \lambda_6) &\geq \lambda_5 + \lambda_6 + v(\lambda_2 + \lambda_3)\end{aligned}$$

These two inequalities express the fact that y beats x in the triple y beats z in the triple.

The following Theorem characterizes the set of choice observations that can be explained using a voting rule.

Theorem 1 *Suppose that $|X| = 3$. For any choice function c there exists a full measure set of explanations $E(c)$.*

The content of Theorem 1 is that this theory of choice is rich enough to explain any choice function on three-element sets. The result can be viewed as a limitation of the theory as it also implies that it cannot be rejected based on choice data from three-element sets. On the other hand, as mentioned earlier, the psychology evidence suggests that in fact all choice patterns can be observed. Finally, the specific structure of a choice problem may impose constraints on the motivations and conflicts that are feasible. For example, we may know from external evidence that certain "reasons" or motivations are simply absent in a given choice problem. Alternatively we may deem certain preferences to be a priori unreasonable and we may restrict the theory to give them a zero weight in any explanation. Such reduction of the set of possible orderings could result a greatly reduced set of explanations, indeed it can produce situations in which some choice patterns have no explanation within the theory as illustrated by the example below. Without such external restrictions however, the theorem above tells us that no behavior pattern can be ruled out a priori.

Example 6 *Each alternative $a = (m_a, e_a)$, $a \in X$, is a job opportunity with two attributes. To fix ideas m_a is a monetary compensation for a and e_a is an "ethical reward" measured in a numerical scale. Alternatives are such that $m_x > m_y > m_z$ and $e_z > e_y > e_z$.*[18] *There are two motivations that rank according to one attribute alone. Using the above notation, the motivation ranking according to the monetary compensation is π_1 and the*

[18] Thus x could be a lucrative job that involves criminal activities, z a low paid job helping the poor, and y an academic job.

motivation ranking according to the ethical reward is π_6. Thus, the domain of populations is restricted by $\lambda_1 + \lambda_6 = 1$. It is straightforward to check that if c display either cyclic choice or third-place choice then $E(c)$ is either empty or has measure zero.

Theorem 1 is a consequence of the following Lemma, a special case of the results presented in the next section.

Proposition 1 *Suppose that $|X| = 3$. If $v \in V_s$ is a voting rule other than the Borda count ($v \neq \frac{1}{2}$) then for any choice function c and any there exists a full measure set $E(c, v)$ of explanations of c based on v. If v is the Borda count ($v = \frac{1}{2}$) then for any choice function c that does not display third-place choice there exists a full measure set $E(c, v)$ of explanations of c based on v.*

The Lemma says that, excluding the Borda count, for any voting rule v and any given choice function c it possible to find populations of motivations λ such that (λ, v) is an explanation of c. The fact that third-place choice is ruled out by the Borda count illustrates that restricting the set of voting rules can also reduce the behaviors explained by the theory (see also Proposition 2). We note that each set $E(c, v)$ is a polytope defined by inequalities just like those in (6a)-(6e). Thus, for each v, the simplex of motivations Δ^6 is cut up into 24 pieces each of which is a polyhedron within which the choice rule is constant.[19] All of these polyhedra are pointed cones that touch at one point, the profile that puts equal weight to each ordering in Π. Each of the polyhedra that contains a vertex of the simplex -a pure motivation- is associated with seemingly rational choice.

5 The General Case

The results presented for the three-alternative case generalize for any number of alternatives.

Theorem 2 *For any choice function c there exists a full measure set of explanations $E(c)$.*

[19] Any element of the polytope $E(c, v)$ is a convex combination of its vertices. For $|X| = 3$, the vertices of $E(c, v)$ for each type of choice function c and arbitrary voting rule v are available at the authors web pages.

The Theorem follows from the following Lemma, which is originally due to Saari (1989, 2001).[20]

Lemma 1 *For any function c and almost any voting rule v there exists a full measure set of explanations $E(c, v)$ based on v.*

A proof is provided in the Appendix. We outline the main argument of the proof. Observe that each voting rule v defines a linear scoring map $S : \Delta^{\Pi} \to \Sigma$ where Δ^{Π} is the set of populations and $\Sigma = \underset{A \in \mathcal{A}}{\times} \Sigma_A$ is a space of all possible scores, with $\Sigma_A \subset \mathbb{R}^{|A|}$. Indeed, for each subset of alternatives $A \in \mathcal{A}$, a population $\lambda \in \Delta^{\Pi}$ produces a vector of scores $s(A, \lambda) \in \Sigma_A$ that has as components the score of each alternative in A. Since these scores are linear in λ, so is $s(A, \lambda)$. The scoring map S is a stack of all of these vectors. The result is established by showing that, except for a lower dimensional set of voting rules, the map S is onto: for any possible stack of scores $\sigma \in \Sigma$ we can find a population $\lambda \in \Delta^{\Pi}$ such that $S(\lambda) = \sigma$. This means that we can generate any possible scores for each alternative and each subset and, thus, any choice function by varying the population of motivations. A key insight for the result is that, for a generic scoring rule, populations that achieve the same voting results (choices) for a subset of the domain $\mathcal{B} \subset \mathcal{A}$, can achieve arbitrarily different results for subsets that are not in $\mathcal{B}$. In particular, the voting results obtained for pairs of alternatives place no constraint on the results for larger subsets. Scoring rules have just enough degrees of freedom to make this possible.

Definition 5 *A voting rule $v = (\gamma_2, ..., \gamma_n)$ is said to be a generalized Borda rule if the k-alternative voting vector γ_k, $k \in \{2, ..., n\}$, satisfies*

$$\gamma_k^j = \frac{j-1}{k-1}\gamma_{k-1}^{j-1} + \frac{k-j}{k-1}\gamma_{k-1}^{j}$$

for $j \in \{2, ..., k-1\}$.

Generalized Borda rules define a lower dimensional family of scoring rules. These rules satisfy a form of consistency: Fix a set $A \in \mathcal{A}$, $a \in$

[20] The authors thank J.P. Benoit for pointing this out. The proof presented herein is based on an argument similar to Saari [1989].

A, and let $K(a, A)$ be the collection of $|A| - 1$ subsets of A that contain a. It is easy to check that for any generalized Borda rule $s(a, A, \lambda) = \frac{1}{|A|-1} \sum_{B \in K(a,A)} s(a, B, \lambda)$ for any population λ. That is, the score of a from A is just the average score of a across subsets in $K(a, A)$. But then, $s(a, B, \lambda)$ is the average score of a across the $|A| - 2$ subsets in $K(a, B)$, and so on. Continuing with this recursion we conclude that for generalized Borda rules, the scores for any subset are fully determined by the scores obtained for pairs of alternatives. As a consequence, it will not be possible to explain all choice rules using generalized Borda rules and we have the following Proposition.

Definition 6 *Fix a choice function c and $A \in \mathcal{A}$. Two alternatives $x_f, x_l \in A$ are pairwise separated by A for c if (i) $c(\{x_f, x\}) = x_f$ for all $x \in A \backslash \{x_f\}$ and (ii) $c(\{x_l, x\}) = x$ for all $x \in A \backslash \{x_l\}$.*

Condition (i) says that c has an alternative x_f that is a "Condorcet winner" among those in A and (ii) says that x_l is a "Condorcet loser" in A. In particular, this means that each of the scores of x_f in pairwise contests is higher than the scores obtained by x_l in pairwise contests.

Proposition 2 *Suppose that a choice function c is such that there exists a subset of three or more alternatives $A \in \mathcal{A}$ such that (i) $x_f, x_l \in A$ are pairwise separated by A for c and (ii) $c(A) = x_l$. If v is a generalized Borda rule then there exists no explanation of c based on v, i.e., $E(c, v) = \emptyset$.*

Note that if $|X| = 3$, (i) and (ii) is equivalent to saying that c displays third-place choice. Condition (ii) says that even though x_l is a Condorcet loser in A it chosen when all alternatives in A are chosen. Proposition 2 illustrates the fact that by restricting the family of aggregation procedures, choice functions that exhibit certain types of irrationality will be ruled out.

6 Ordinal Welfare Inferences

This section presents the paper's main normative conclusions. We ask whether a limited form of behavior consistency leads to a similar welfare inference than in the case of the rational choice model.

6.1 Pairwise Coherence

In the rational choice model choice and welfare are perfectly aligned: choices are consistent and if choice data reveals that x is always chosen over y this is because the decision-maker's welfare is higher under x than it is under y. Our welfare method allows for choice patterns that are not compatible with rational choice. If the choice data is such that x is always chosen over y, can we infer that that welfare increases if the available set changes from a set A having $y = c(A)$ to a set B having $x = c(B)$? This motivates a key concept of limited consistency:

Definition 7 *A choice function c is said to be pairwise coherent or simply pairwise coherent with respect to $(x, y) \in X^2$ if for any A that contains both x and y we have that $c(A) \in \{x, y\} \Rightarrow c(A) = x$.*

Pairwise coherence is a version of the consistency axioms of rational choice theory valid for particular pair rather than all pairs.[21] In the case of three-alternatives, there two type of choice functions that exhibit pairwise coherence. Using the notation introduced in section 3.1, the seemingly rational choice is pairwise coherent w.r.t two pairs: (x^r, x^s) and (x^s, x^t). The second type are choice functions that display cyclic choice. If c displays cyclic choice and is given by (5), then c is pairwise coherent w.r.t (x, y).

As before we use $F = (A, B) \in \mathcal{A} \times \mathcal{A}$ to designate a change in the available set from A to B and $\Delta W^+(F)$ to denote the set of choice functions that admit a positive welfare inference for F (see definition 3).

Theorem 3 *Let $A, B \in \mathcal{A}$ with $y = c(A)$ and $x = c(B)$. If c is pairwise coherent w.r.t. (x, y) there exists an explanation $\lambda \in E(c)$ that puts weight exclusively on motivations that prefer x over y. In particular, if $F = (A, B)$ then $c \in \Delta W^+(F)$.*

Theorem 3 says that if c is such that x is always chosen over y, no matter how crazy the choice pattern c, there is always an explanation of the behavior

[21] The choice functions compatible with rational choice are those satisfying Houtakker's axiom. Let $a, a' \in X$ and $c_{aa'} = \{A \in \mathcal{A} |\ c(A) \in \{a, a'\}\}$. One version of this axiom is as follows: *For any pair of alternatives a and a', either $a \in c(A)$ for all $A \in c_{xy}$ or else $a' \in c(A)$ for all A.* If $c(A)$ is a singleton for all $A \in c_{aa'}$, one can replace "$\in$" with "$=$". Hence, the axiom just says that for c there are no "preference reversals": ignoring indifferences, either a is "revealed preferred" to a' or vice-versa. Pairwise coherence establishes the same for a particular pair $a = x$ and $a' = y$.

that involves motivations which are not conflicted with respect to this pair of alternatives. As a consequence, for this choice function, one can never rule out a welfare improvement if opportunities change from a set where y is chosen to one where x is chosen. It also worth to point out that if c is pairwise coherent w.r.t. (x, y) the for any $\lambda \in E(c)$ the lower bound of on the strength of motivations that prefer x over y is $1/2$. This is an immediate inference of the choice from the pair $c(\{x, y\}) = x$.

The following Proposition is a partial converse to Theorem 3. We denote by $\widetilde{V} \subset V_s$ the non-generic subset of scoring rules that have some voting vector γ_k with equal components, i.e., $\gamma_k^j = \gamma_k^{j+1} = .. = \gamma_k^{j'}$ for some $j, j' \in \{1, ..., k\}$.

Proposition 3 *Suppose that c is not pairwise coherent w.r.t. (x, y). If $v \notin \widetilde{V}$ then there exists no explanation $\lambda \in E(c, v)$.that puts weight exclusively on motivations that prefer one alternative over the other.*

The Proposition says that except for a non-generic voting rules that are characterized by full loss of ordinal information, a choice function c for which there is a "preference reversal" between x and y can only arise from population of motivations that are conflicted with regard to these alternatives.

6.2 Expansion Monotonicity

A second tenet of welfare inference in the rational choice model is that more choice is always better. Indeed, if choice is seemingly rational and thus consistent with the optimization of a preference, the maximal element of this preference must be $c(X)$, the choice when all alternatives are available. The second best element can be inferred by looking at the choice from $X \backslash \{c(X)\}$. If we continue shrinking the set of alternatives by deleting alternatives that are previous chosen, we can infer the entire preference relation. In our case, c need not be consistent with optimization. However, as shown below, the order just described is still important to determine whether expanding the set of opportunities is beneficial.

A change in the opportunities from A to B, $F = (A, B) \in \mathcal{A} \times \mathcal{A}$, is said to be an *expansion* if A is a strict subset of B. We introduce a concept of partial consistency closely related to expansions:

Definition 8 *Let $x_n(c) = c(X)$ and $x_j(c) = c(X \backslash \{x_{j+1}(c), ..., x_n(c)\})$ for $j = 1, ..., n-1$. Let $\succeq_c$ be the order on X defined by $x_j(c) \succeq_c x_k(c)$ if and*

only if $j > k$. A choice function c is said to be expansion-monotonic with respect to $\succeq_c$ if for any expansion $F = (A, B)$ with $A \subset B$ we have that $c(B) \succeq_c c(A)$.

In the case of three alternatives it is easy to check that an expansion-monotonic choice function c must be seemingly rational. With more than three alternatives the set of expansion-monotonic choice functions is larger. An example of a choice function that is expansion-monotonic but not seemingly rational when $X = \{x, y, z, w\}$ has four alternatives is

$$\begin{aligned} c(\{x,y\}) &= y, c(\{x,z\}) = z, c(\{x,w\}) = w, c(\{y,z\}) = y, c(\{y,w\}) = w, c(\{z,w\}) = z \\ c(\{x,y,z\}) &= y, c(\{x,y,w\}) = y, c(\{x,z,w\}) = z, c(\{y,z,w\}) = y \\ c(X) &= x \end{aligned}$$

Throughout the remainder of the section we assume that there exists a welfare functional $\widetilde{W}(\cdot, \lambda) : X \to \mathbb{R}$ that represents welfare for individual with characterized by a population λ. We use $\Delta W(F, c, \lambda) = \widetilde{W}(c(B), \lambda) - \widetilde{W}(c(A), \lambda)$ for the change in welfare corresponding to a change $F = (A, B)$ in available opportunities.

Definition 9 *Fix a welfare functional $\widetilde{W}$. A choice c rule admits positive expansions under $\widetilde{W}$ if there exists $\lambda \in E(c)$ such that for any expansion $F = (A, B)$ we have that $\Delta W(F, c, \lambda) \geq 0$.*

Proposition 4 *Fix any welfare functional $\widetilde{W}$. If c admits positive expansions under $\widetilde{W}$ then c is expansion-monotonic with respect to $\succeq_c$.*

7 Utilitarian Bounds and the Value of Expansions

This section focuses on utilitarian welfare measures. The method described in section 2.3 gives rise to a family of measures, one for each population $\lambda \in E(c)$ consistent with the choice observations. The practical use of this program rests on the ability to compute these measures efficiently. We start by showing how to exploit the structure of the set of explanations $E(c)$ to derive welfare bounds. Next, we illustrate how to use these bounds to determine whether an expansion of the set of opportunities is beneficial or harmful.

7.1 Welfare Bounds and Basic Voting Rules

We describe how to obtain bounds for two sets of welfare measures. The first set of measures is simply $W(A,c,u) = \{W(A,c,\lambda,u)\}_{\lambda \in E(c)}$ as defined by (3). The upper and lower bounds for $W(A,c,u)$ are

$$\begin{aligned} W^{\min}(A,c,u) &= \inf\{W(A,c,\lambda,u) \mid \lambda \in E(c)\} \text{ and} \\ W^{\max}(A,c,u) &= \sup\{W(A,c,\lambda,u) \mid \lambda \in E(c)\}, \end{aligned} \tag{8}$$

the worst and best welfare measures at A.

In the case of utilitarian welfare measures (4) reduces to

$$\Delta W(F,c,\lambda,u) = \sum_{\pi \in \Pi} \Delta u(F,c,\pi)\lambda(\pi), \tag{9}$$

where $\Delta u(F,c,\pi) = u(c(B),\pi) - u(c(A),\pi)$. This is simply the average change in utility for a population $\lambda \in E(c)$ consistent with the observed choice function c. The set of welfare evaluation measures associated to this change in opportunities is $\Delta W(F,c,u) = \{\Delta W(F,c,\lambda,u)\}_{\lambda \in E(c)}$ with upper and lower bounds given by

$$\begin{aligned} \Delta W^{\min}(F,c,u) &= \inf\{\Delta W(F,c,\lambda,u) \mid \lambda \in E(c)\} \text{ and} \\ \Delta W^{\max}(F,c,u) &= \sup\{\Delta W(F,c,\lambda,u) \mid \lambda \in E(c)\}. \end{aligned} \tag{10}$$

Both $W(A,c,\lambda,u)$ and $\Delta W(F,c,\lambda,u)$ are linear functions of λ, so calculating the above bounds amounts to solving a optimization problem with a linear objective of the form

$$Z(c,r) = \min\{r^t\lambda | \lambda \in E(c)\},$$

where $r \in \mathbb{R}^{\Pi}$. It follows that the extremal values will be realized at an extreme point of $E(c)$. However, $E(c)$ is a union of polyhedra which, in general, is not itself a polyhedron nor even a convex set. In principle, finding the extreme points of an arbitrary set can be hard.

For each $j \in \{1,...,k-1\}$, $k \geq 2$, let e_k^j be the $k-$component vector having its first j components equal to one and its last $k-j$ components equal to zero.

Definition 10 *A k-alternative voting vector $\gamma_k \in [0,1]^k$ is said to be* basic *if $\gamma_k = e_k^j$ for some $j \in \{1,...,k-1\}$. A voting rule $v = (\gamma_2,...,\gamma_n)$ is said to be basic if for each $k \in \{2,...,|X|\}$ vector γ_k is a basic voting vector. The set of basic voting rules is denoted by V^{basic}.*

It is easy to verify that if $n = |X|$, there are $(n-1)!$ basic voting rules. If $|X| = 3$ the only two basic voting rules are Plurality and Antiplurality.

Proposition 5 *Fix $r \in \mathbb{R}^{\Pi}$ and a choice rule c. Let $Z(c,r,v) = \min\{r^t\lambda | \lambda \in E(c,v)\}$. Then $Z(c,r) = Z(c,r,\widehat{v})$ for some basic voting rule $\widehat{v}$. In particular, $Z(c,r) = \min\{Z(c,r,v) | v \in V^{basic}\}$.*

Observe that $Z(c,r,v) = \min\{r^t\lambda | \lambda \in E(c,v)\}$ is a standard linear program (LP) as $E(c,v)$ is a polyhedron. Thus, Proposition 5 provides a simple procedure to compute utilitarian bounds: (1) Compute the corresponding lower (upper) bound for each basic voting rule by solving a LP; (2) Use the smallest (largest) of these values.

Corollary 1 *Suppose that $|X| = 3$. Each of the bounds defined by (8) and (10) over all explanations is attained by a population which is either a vertex of the polytope $E(c,P)$ or a vertex of $E(c, Anti - P)$.*

To compute an upper bound it suffices to solve two LP, one where the feasible set is $E(c,P)$ and one in which the feasible set is $E(c, Anti - P)$. The upper bound is simply the highest of the values associated to these LP. Similarly, lower bounds are obtained by solving two linear minimization programs, one where the feasible set is $E(c,P)$ and one in which the feasible set is $E(c, Anti - P)$. The lower bound is the minimum of the values associated to these LP.

7.2 Surely Beneficial and Surely Harmful Expansions

Consider an expansion $F = (A,B)$, where $A \subset B$. If we observe a choice pattern c, what can we say about the set of welfare measures associated to the expansion $\Delta W(F,c,u)$? Our goal is to identify restrictions on the cardinalization profile that allow to determine whether an expansion is surely beneficial or surely harmful.

Definition 11 *Fix a choice rule c and an expansion $F = (A,B) \in \mathcal{A} \times \mathcal{A}$. If the expansion changes the choice, $c(A) \neq c(B)$, the choice rule is said to be* choice-varying *at F. The set of choice-varying expansions for c is denoted by $\mathcal{F}(c)$.*

If an expansion does not change the choice then the welfare change is trivially zero. We restrict attention to expansions in $\mathcal{F}(c)$, i.e., those associated to changes in the choice.

Since the welfare measure is an average across the population, these restrictions depend crucially on two variables, the relative strength of motivations favored by the expansion vis a vis those hurt by it, and the magnitude of the gains and losses associated to each group of motivations.

Given a choice rule c and an expansion $F = (A, B) \in \mathcal{F}(c)$, let $\Pi^+(F, c) = \{\pi \in \Pi \mid c(B)\pi(A)\}$ be the set of motivations that gain from the expansion. The set of motivations that are hurt by the expansion is $\Pi^-(F, c) = \Pi \setminus \Pi^+(F, c)$. For each $\lambda \in E(c)$ we define

$$\mu^+(F, c, \lambda) = \sum_{\pi \in \Pi^+(F,c)} \lambda(\pi)$$

and

$$\mu^-(F, c, \lambda) = \sum_{\pi \in \Pi^-(F,c)} \lambda(\pi) = 1 - \mu^+(F, c, \lambda),$$

the strength of motivations favored respectively hurt by this expansion. The ratio

$$\rho(F, c, \lambda) = \frac{\mu^+(F, c, \lambda)}{1 - \mu^+(F, c, \lambda)}$$

measures the relative strength of these motivations. For a symmetric cardinalization profile in which all motivations have the same utilities, this number is sufficient to determine whether an expansion is beneficial or harmful at λ. However, we may not be able to compute this measure as λ is not directly observed. Knowing that $\lambda \in E(c)$ it is possible to estimate the upper and lower bounds given by

$$\underline{\rho}(F, c) = \inf_{\lambda \in E(c)} \rho(F, c, \lambda) = \frac{\underline{\mu}^+(F, c)}{1 - \underline{\mu}^+(F, c)}$$

and

$$\overline{\rho}(F, c) = \sup_{\lambda \in E(c)} \rho(F, c, \lambda) = \frac{\overline{\mu}^+(F, c)}{1 - \overline{\mu}^+(F, c)}.$$

Here $\underline{\mu}^+(F, c) = \min_{\lambda \in E(c)} \mu^+(F, c, \lambda)$ and $\overline{\mu}^+(F, c) = \max_{\lambda \in E(c)} \mu^+(F, c, \lambda)$. Since $\mu^+(F, c, \lambda)$ is linear in λ, Proposition 5 and Corollary 1 can be used to calculate these numbers.

Finally, we introduce notation for bounds on the gains and losses associated to an expansion $F \in \mathcal{F}(c)$ for a fixed cardinalization u. Let $g(u,F,c) = \min_{\pi \in \Pi^+(F,c)} |\Delta u(F,X,c,\pi)|$ respectively $G(u,F,c) = \max_{\pi \in \Pi^+(F,c)} |\Delta u(F,c,\pi)|$ be the smallest respectively largest gain across motivations that benefit from the expansion. Let $l(u,F,c) = \min_{\pi \in \Pi^-(F,c)} |\Delta u(F,c,\pi)|$ respectively $L(u,F,c) = \max_{\pi \in \Pi^-(F,c)} |\Delta u(F,c,\pi)|$ be the smallest respectively largest loss across motivations hurt by the expansion.

Theorem 4 *Fix a cardinalization u, a choice rule c, and a choice-varying expansion $F \in \mathcal{F}(c)$.*

(i) If $\dfrac{L(u,F,c)}{g(u,F,c)} < \underline{\rho}(F,c)$ *then* $\Delta W^{\min}(F,c,u) > 0$;

(ii) If $\dfrac{l(u,F,c)}{G(u,F,c)} > \overline{\rho}(F,c)$ *then* $\Delta W^{\max}(F,c,u) < 0$.

Theorem 4 provides two "likelihood ratio test" conditions.[22] The left hand side of the inequality in (i) is an upper bound on the loss of a motivation hurt by the expansion relative to the gain from a motivation that benefits from it. The right hand side $\underline{\rho}$ is a lower bound on the ratio of the strengths of motivations that benefit and those that lose from the expansion. These ratios are independent of u. Lemma 2 below provides these bounds for each possible c and $F \in \mathcal{F}(c)$ in the case $|X| = 3$. The condition says that if the relative loss is small relative to the strengths ratio an expansion is surely beneficial for that cardinalization. The second condition compares a lower bound of the relative loss with an upper bound $\overline{\rho}$ of the strengths ratio. If satisfied an expansion is surely detrimental for that cardinalization. The example below illustrates that these bounds are tight.

Definition 12 *Fix a choice rule c. An expansion $F = (A,B) \in \mathcal{F}(c)$ is said to be pairwise coherent if c is pairwise coherent w.r.t. $(c(B), c(A))$.*

Observe that, from Theorem 3, if F is a pairwise coherent then $\overline{\rho}(F,c) = +\infty$. Hence, for a pairwise coherent expansion the hypothesis of (ii) will not be satisfied for any cardinalization u. This means that pairwise coherent

[22] In classical decision theory, the optimal decision to accept or reject a hypothesis is often expressed as condition based on a critical value for the likelihood ratio.

expansions are always beneficial at some explanation, something we already knew from Theorem 3.

Using Corollary 1 we obtain the following:

Lemma 2 *Suppose that* $|X| = 3$. *Fix a choice rule* c *and a choice-varying expansion* $F \in \mathcal{F}(c)$.

(i) If c *is seemingly rational* $\underline{\rho}(F, c) = 1$ *and* $\overline{\rho}(F, c) = +\infty$;

(ii) If c *displays second-place choice* $\underline{\rho}(F, c) = 0$ *and* $\overline{\rho}(F, c) = 1$;

(iii) If c *displays third-place choice* $\underline{\rho}(F, c) = \frac{1}{3}$ *and* $\overline{\rho}(F, c) = 1$;

(iv) If c *displays cyclic choice* $\underline{\rho}(F, c) = 1$ *and* $\overline{\rho}(F, c) = +\infty$ *if* F *is a pairwise coherent expansion and* $\underline{\rho}(F, c) = \frac{1}{3}$ *and* $\overline{\rho}(F, c) = 1$ *if* F *is not pairwise coherent.*

Larger values of ρ are associated with a higher presence of motivations that benefit from the expansion. The strength of motivations that benefit from expansions is highest for seemingly rational choice rules. Cyclic choice seems to dominate second-place and third-place choice functions in this dimension. Interestingly, the upper and lower bounds associated to the pairwise coherent expansion for cyclic choice are exactly the same than those associated to the expansion for seemingly rational choice (which is also pairwise coherent).

Example 7 *Harmful Expansion with Seemingly Rational choice*

Suppose that $X = \{x, y, z\}$. Let c be the seemingly rational choice function described by (6) and $A = \{y, z\}$. Consider the expansion $F = (A, X)$ and observe that $c(A) = y$ and $c(X) = x$. The set of motivations that gain is $\Pi^+(F, c) = \{\pi_1, \pi_2, \pi_5\}$ and $\Pi^-(F, c) = \{\pi_3, \pi_4, \pi_6\}$.

Consider a cardinalization u such that $\Delta u(F, c, \pi) = 1$ for all $\pi \in \Pi^+(F, c)$ and $\Delta u(F, c, \pi) = -(1 + \alpha)$ for all $\pi \in \Pi^-(F, c)$, $\alpha > 0$. Thus, $g = G = 1$ and $L = l = 1 + \alpha$. From Lemma 2, $\underline{\rho}(F, c) = 1$. Thus, this cardinalization violates the hypothesis of (i) as $\alpha > 0$. We show that at some explanation $\Delta W < 0$. Indeed, for a small $\epsilon > 0$, consider the population λ defined by

$$\lambda_1 = \tfrac{1}{2} + \epsilon,\ \lambda_2 = \epsilon,\ \lambda_4 = \tfrac{1}{2} - 2\epsilon.$$

It is easy to verify that λ and any voting rule v explains c. The welfare change for this profile is

$$\Delta W = -\tfrac{1}{2}\alpha + \epsilon(2+\alpha).$$

Clearly, for any α there exists an ϵ small enough such that $\Delta W < 0$.

The example illustrates two issues. First, the hypothesis of (i) is tight: a small departure from the condition (α is arbitrary) can reverse the conclusion. The same is true for the hypothesis of (ii). Secondly, it also shows that expansions can be harmful even if observed behavior is consistent with rationality.

8 Conclusion

We have modeled potentially non-rational choice as a conflict between simultaneously held motivations with possibly different strengths. We assume, following the long-standing tradition of revealed preference theory, that the individual's choice is observed and we seek to describe it "as if" it arose from the aggregation of conflicting preferences. In this paper we have restricted ourselves to aggregation rules that are purely ordinal, taking the form of a scoring rule.

The task of economic analysis is to determine which conflicting preferences could give rise to the observed choice function. Once these preferences are known, welfare analysis can use them to evaluate the efficacy of any given change in the set of alternatives that is available.

We have given a general method under which this program can be carried out and we have applied it to the case in which there are only three possible alternatives. We find all the possible explanations of a choice rule. The reason for needing all the explanations is that we are interested in developing welfare bounds and it is therefore necessary to look for best and worst case evaluations within the set of all explanations.

While the three alternative case may seem quite restrictive, it is widely documented in psychological experiments. It forms a good starting point for testing our method. Indeed our method bears out some of the main experimental observations: There are four possible patterns of choice; all have been experimentally observed. Our theory shows that all are in fact possible as the result of the aggregation of conflicting motivations by voting methods that are scoring rules.

The principal benefit from using our method to find the set of underlying preferences is that we can use these preferences to measure welfare changes. In the case of an expansion of the available alternatives, we show that there are some changes that are surely beneficial while others may or may not be beneficial, and we demonstrate the sensitivity of this result choice function that has been observed and to the cardinalization of utilities that represent the intensity of preference within each motivation.

In future work we will try to allow for conflicting motivations to be aggregated by rules that incorporate some of the cardinal information available. We will also look for criteria under which some explanations can be dropped and others can be highlighted. The aim in all these endeavors will be to provide sharper welfare bounds than those available from the method in the present paper alone.

A Example

We let $\varrho = 0.3$, meaning that each period of delay in consumption reduces the utility of consumption to 0.74 of its prior value. It is easiest to take the logarithm of the utility, obtaining $h_i(a) = \phi_{ia} - \varrho t_a$, where $\phi_{ia} = \ln \tilde{u}_i(m_a)$. We want to insure that the preferences captured by the functions u_i do in fact rank the three pairs (x_a, t_a) in all six of the possible orders. We also need to respect the monotonicity of the $\tilde{u}_i$, which requires $\phi_{ix} < \phi_{iy} < \phi_{iz}$ for $i = 1, ..., 6$. Consider the following values for the parameters ϕ_{ia} :

$i \setminus a$	x	y	z
1	0.20	0.45	0.70
2	0.20	0.40	0.75
3	0.15	0.50	0.70
4	0.10	0.50	0.75
5	0.10	0.45	0.80
6	0.15	0.40	0.80

Using these ϕ_{ia} it can be verified that

$$\begin{aligned} h_1(x) &> h_1(y) > h_1(z) \\ h_2(x) &> h_2(z) > h_2(y) \\ h_3(y) &> h_3(z) > h_3(x) \\ h_4(y) &> h_4(x) > h_4(z) \\ h_5(z) &> h_5(x) > h_5(y) \\ h_6(z) &> h_6(y) > h_6(x). \end{aligned}$$

Thus, each motivation u_i has a different ranking over the alternatives. It is also convenient that the realized values $h_i(a) = 0.2, 0.15$ and 0.1, depending on whether a is first, second, or third in the order specified by preference i. It can also be verified that each $\tilde{u}_j$ is risk averse over the three possible payoffs, 10, 15 and 35.

B Proofs

B.1 Explanations

Proof of Lemma 1

We omit reference to the voting rule v which is assumed to be fixed. Let $n = |X|$ and $Q = |\mathcal{A}| = 2^n - n - 1$ be the number of subsets of two or more alternatives. The number of orderings over X is $|\Pi| = n!$.

We start by noting that any explanation $\lambda \in \Delta^\Pi$ can be decomposed as $\lambda = \delta + \frac{e_\Pi}{|\Pi|}$, where $e_\Pi \in \mathbb{R}^\Pi$ is a vector of ones. Hence, $\frac{e_\Pi}{|\Pi|} \in \Delta^\Pi$ is the explanation that assigns equal weight to each preference in Π and δ is the deviation from this explanation. Since λ is in the simplex, the deviation vector δ satisfies $e_\Pi^T \delta = 0$. Note also that, for any explanation λ and any subset A, adding the scores across alternatives in A gives the total number of points $W_A = \sum_{j=1}^{|A|} \gamma_{|A|}^j$, which is independently of λ. This means that, for each subset and each for deviation δ, the scores of alternatives add up to zero . For any subset A, the profile $\frac{e_\Pi}{|\Pi|}$ yields the same score for each alternative $a \in A$. The result is established if we show that, for any c it is always possible to find a deviation δ that breaks these ties as specified by c. This is done by showing that we can achieve any possible scores for each subset as we vary the deviation $\delta \in \Delta^* = \{x \in \mathbb{R}^\Pi \mid e_\Pi^T x = 0\}$.

We need to introduce some notation. Each voting rule induces a linear scoring map $S : \mathbb{R}^\Pi \to E$, where $E = \underset{A \in \mathcal{A}}{\times} \mathbb{R}^{|A|}$. To each vector $x \in \mathbb{R}^\Pi$ we associate the score of alternative a from subset $A \in \mathcal{A}$

$$s(a, A, x) = (\gamma_A)^T P^{aA} x \tag{11}$$

where γ_A is the $|A| \times 1$ voting vector for a subset of size $|A|$ and P^{aA} is a $|A| \times |\Pi|$ matrix. The $i\pi$ element of this matrix is $P_{i\pi}^{aA} = \delta_{ir(a,A,\pi)}$ where $r(a, A, \pi)$ is the ranking of alternative a in A for motivation π. The vector of scores $s(A, x)$ for subset A is an element of $\mathbb{R}^{|A|}$ and the stack of all of these vectors is $S(x)$, an element of E. Observe that the dimension of E is $M = \sum_{k=2}^n \binom{n}{k} k$ as there are $\binom{n}{k}$ subsets of $k \geq 2$ alternatives ($M = 2^{n-1}n - Q$).

The domain $\mathbb{R}^\Pi$ is a linear space that can be decomposed into two subspaces $\Delta^* = \{x \in \mathbb{R}^\Pi \mid e_\Pi^T x = 0\}$ and the space orthogonal to it, $\Delta_\perp^*$, which is spanned by the vector e_Π. Clearly, $\dim(\Delta^*) = \Pi - 1$ and $\dim(\Delta^\perp) = 1$ and each $x \in \mathbb{R}^\Pi$ can be written as $x = x_\Delta + x_\perp$, with $x_\Delta \in \Delta^*$ and $x_\perp \in \Delta^\perp$. Similarly, let $\Sigma = \{s \in \mathbb{R}^A \mid e_A^T s = 0\}$ and $\Sigma = \underset{A \in \mathcal{A}}{\times} \Sigma_A$. Note that Σ is a set of scores that sum to zero for each subset and, as argued earlier, $S(\Delta^*) \subseteq \Sigma$. The codomain E can be decomposed into Σ and the space $\Sigma^\perp$ orthogonal to it. Since Σ is defined by one constraint for each of the Q subsets in $\mathcal{A}$, we have that $\dim(\Sigma) = M - Q$ and $\dim(\Sigma^\perp) = Q$.

To establish the Lemma it suffices to show that for any vector of scores $s \in \Sigma$ there is a deviation vector $\delta \in \Delta^*$ such that $s = S(\delta)$, that is, $S(\Delta^*) = \Sigma$. We proceed in two steps.

Step 1: If $\dim(\ker(S)) = Q - 1$ then $S(\Delta^*) = \Sigma$.

Since $S(\Delta^*) \subseteq \Sigma$, we have that $S(\Delta^*) = \Sigma$ if $\dim(S(\Delta^*)) = \dim(\Sigma)$. Now, $\dim(\Sigma) = M - Q$ so the previous holds if and only if $\dim(S(\Delta^*)) = M - Q$. We show that the latter holds if and only if if $\dim(\ker(S)) = Q - 1$.

By the dimension decomposition theorem we know that

$$\dim(S(\mathbb{R}^{\Pi})) + \dim(\ker(S)) = \dim(E) = M.$$

We also know that $\dim(S(\Delta^{\perp})) = 1$ as $\dim(\Delta^{\perp}) = 1$ and, by construction,

$$\dim(S(\mathbb{R}^{\Pi})) = \dim(S(\Delta^{\perp})) + \dim(S(\Delta^*)).$$

Combining the previous we have

$$\dim(S(\Delta^*)) = M - 1 - \dim(\ker(S)),$$

and the conclusion follows.

Step 2: $\dim(\ker(S)) = Q - 1$

Let L be the matrix associated to S. We establish Step 2 by showing that for any $\alpha \in E$ satisfying $L^T\alpha = 0$ then $\alpha_{aA} = \gamma_A$ for each $a \in A$, and $\sum_{A \in \mathcal{A}} \gamma_A = 0$. That is, the nullspace of L^T is defined by one parameter per subset and these parameters must sum to zero, which means that it has dimension $Q-1$. We use induction $k = \{2, ..., n\}$ over the size of each subset.

We introduce some notation. Let $\mathcal{A}_k = \{A \in \mathcal{A} |\ |A| \geq k\}$ be the family containing all subsets of k or more alternatives, $k = \{2, ..., n\}$. Clearly, $\mathcal{A}_2 = \mathcal{A}$ and $\mathcal{A}_k$ is decreasing. For each each $\lambda \in \Delta^{\Pi}$ define

$$h_k(\lambda) = \sum_{a \in A, A \in \mathcal{A}_k} s(a, A, \lambda)\alpha_{aA}.$$

It is straightforward to check that if α satisfies $L^T\alpha = 0$ we have that

$$\sum_{a \in A, A \in \mathcal{A}} s(a, A, \pi)\alpha_{aA} = 0 \text{ for all } \pi \in \Pi,$$

which by the linearity of the scores with respect to λ is equivalent to

$$h_2(\lambda) = \sum_{a \in A, A \in \mathcal{A}} s(a, A, \lambda)\alpha_{aA} = 0 \text{ for all } \lambda \in \Delta^{\Pi}. \tag{12}$$

We are now in position of showing our initial induction step. Let $B = \{b_1, b_2\}$ be an arbitrary subset of size $k = 2$. From Lemma 3, there exist λ and $\widehat{\lambda}$ such that $s(a, A, \lambda) = s(a, A, \widehat{\lambda})$ for all $a \in A$, $A \in \mathcal{A}_2 \backslash \{B\}$. Further, $s(b_1, B, \lambda) = s(b_2, B, \widehat{\lambda}) = 1$ and $s(b_2, B, \lambda) = s(b_1, B, \widehat{\lambda}) = 0$. Combining this with (12) we get

$$h_2(\lambda) - h_2(\widehat{\lambda}) = \alpha_{b_1 B} - \alpha_{b_2 B} = 0$$

or $\alpha_{b_1 B} = \alpha_{b_2 B} = \gamma_B$, which completes the initial step.

Our induction hypothesis is that $\alpha_{aA} = \gamma_A$ for each $a \in A$ and $|A| = k - 1$. We need to show that the previous is also satisfied for subsets of k alternatives. Indeed,

$$h_2(\lambda) = \sum_{a \in A, A \in \mathcal{A}_2 \backslash \mathcal{A}_k} s(a, A, \lambda) \alpha_{aA} + h_k(\lambda)$$

and, using the induction hypothesis, we have that

$$\sum_{a \in A, A \in \mathcal{A}_2 \backslash \mathcal{A}_k} s(a, A, \lambda) \alpha_{aA} = \sum_{A \in \mathcal{A}_2 \backslash \mathcal{A}_k} \gamma_A \sum_{a \in A} s(a, A, \lambda) = \sum_{A \in \mathcal{A}_2 \backslash \mathcal{A}_k} \gamma_A W_A,$$

where $W_A = \sum_{j=1}^{|A|} \gamma_{|A|}^j$ which is independent of λ. Hence, using (12) we conclude that

$$h_k(\lambda) = c_k \text{ for all } \lambda \in \Delta^{\Pi}, \tag{13}$$

where c_k is a constant independent of $\lambda \in \Delta^{\Pi}$. Invoking Lemma 3 again, there exist $\lambda', \lambda'' \in \Delta^{\Pi}$ λ' and λ'' such that $s(a, A, \lambda') = s(a, A, \lambda'')$ for all $a \in A$, $A \in \mathcal{A}_k \backslash \{B\}$. Further, $s(b_1, B, \lambda') = s(b_2, B, \lambda'') = 1$ and $s(b_2, B, \lambda') = s(b_1, B, \lambda') = 0$. Using this and (13) we conclude that, for any B, and $\{b_1, b_2\} \subset B$, with $|B| = k$, we have that

$$h_k(\lambda') - h_k(\lambda'') = \alpha_{b_1 B} - \alpha_{b_2 B} = 0.$$

This shows that $\alpha_{b_1 B} = \alpha_{b_2 B}$. Since b_1 and b_2 are arbitrary elements of B and this is an arbitrary subset of size k, we conclude that $\alpha_{bB} = \gamma_B$ for any such set. The induction is complete. ■

Lemma 3 *For almost any voting rule v, any subset of alternatives $B \in \mathcal{A}$, and any pair $\{b_1, b_2\} \subseteq B$, there exist a pair of profiles λ and $\widehat{\lambda}$ such that $s(a, A, \lambda) \neq s(a, A, \widehat{\lambda})$ for $a \in A$, and $A \in \mathcal{A}$ with $|A| \geq |B|$ if and only if $A = B$ and $a \in \{b_1, b_2\}$.*

Proof. We provide a complete argument for $v \in \{Plurality, Anti-Plurality\}$. Fix $B \in \mathcal{A}$ and a pair $\{b_1, b_2\} \subseteq B$. We construct preferences π and $\widehat{\pi}$ that give rise to identical scores for all alternatives in subsets other than B having $|B|$ or more alternatives. We also show that $s(a, B, \pi) = s(a, B, \widehat{\pi})$ unless $a \in \{b_1, b_2\}$. We later explain how to extend the argument for a general v.

Consider π to be a preference such that $b_1 \pi b_2$ and there are there are no elements of X ranked between alternatives b_1 and b_2. Let $\widehat{\pi}$ be identical to π except for the fact that $b_2 \pi b_1$, i.e., b_1 and b_2 have been permuted. Observe that π and $\widehat{\pi}$ induce the same order over any subset of alternatives that does not contain the pair $\{b_1, b_2\}$. In particular, for any A such that $\{b_1, b_2\} \nsubseteq A$, we have that $s(a, A, \pi) = s(a, A, \widehat{\pi})$ for each $a \in A$. Throughout the proof we consider π and $\widehat{\pi}$ as above.

Suppose that $v = Plurality$. Let $M(A, \widetilde{\pi})$ denote the maximal element in A for preference $\widetilde{\pi} \in \Pi$ and observe that, for Plurality, this is the only alternative that receives a score of one (all others get a zero score). Consider π and $\widehat{\pi}$ satisfying two additional properties: (i) $a \pi b$ and $a \widehat{\pi} b$ for any $a \in X \backslash B$, $b \in B$; and (ii) In B, b_1 is ranked first by π and b_2 is ranked first by $\widehat{\pi}$. Now, since any subset $A \neq B$ having $|A| \geq |B|$ contains elements in $X \backslash B$, we have that $M(A, \pi) = M(A, \widehat{\pi}) \in X \backslash B$ as elements in B are ranked at the bottom for both π and $\widehat{\pi}$ (by (i)) and these preferences are identical except for the permutation of b_1 and b_2. We conclude that $s(a, A, \pi) = s(a, A, \widehat{\pi})$ for each $a \in A$, $A \neq B$, and $|A| \geq |B|$. Finally (ii) implies that $s(b_1, B, \pi) = s(b_2, B, \widehat{\pi}) = 1$ and $s(b_1, B, \widehat{\pi}) = s(b_2, B, \pi) = 0$.

Suppose instead that $v = Anti-Plurality$. Let $m(A, \widetilde{\pi})$ denote the worse element in A for preference $\widetilde{\pi} \in \Pi$ and observe that, for Anti-Plurality, this is the only alternative that receives a score of zero (all others get a score of one). Consider π and $\widehat{\pi}$ satisfying two properties: (i) $b \pi a$ and $b \widehat{\pi} a$ for any $a \in X \backslash B$, $b \in B$; and (ii) In B, b_2 is ranked last by π and b_1 is ranked last by $\widehat{\pi}$. Any subset $A \neq B$ having $|A| \geq |B|$ contains elements in $X \backslash B$. Thus, from (i), $m(A, \pi) = m(A, \widehat{\pi}) \in X \backslash B$. It follows that $s(a, A, \pi) = s(a, A, \widehat{\pi})$ for each $a \in A$, $A \neq B$, and $|A| \geq |B|$. It is straightforward to check that (ii) implies that $s(b_1, B, \pi) = s(b_2, B, \widehat{\pi}) = 1$ and $s(b_1, B, \widehat{\pi}) = s(b_2, B, \pi) = 0$.

To extend the argument for a general rule v, let $\Pi_{12} = \{\pi | b_1 \pi b_2\}$ and $\Pi_{21} = \{\pi | b_2 \pi b_1\}$. Define $f : \Pi_{12} \to \Pi_{21}$ to be the map that assigns to each $\pi \in \Pi_{12}$ the preference $\widehat{\pi} = f(\pi) \in \Pi_{21}$ that ranks all elements the same as π except that b_1 and b_2 have been permuted in the order. As in the cases above, π and $\widehat{\pi}$ induce the same order over any subset of alternatives that does not contain the pair $\{b_1, b_2\}$. The populations λ and $\widehat{\lambda}$ in the statement of the

lemma are such that $\lambda(\pi) = \widehat{\lambda}(f(\pi))$, from which $s(a, A, \pi) = s(a, A, \widehat{\pi})$ for each $a \in A$, as long as $\{b_1, b_2\} \nsubseteq A$. The specific weights $\lambda(\pi)$ are constructed to ensure that $s(b_1, B, \lambda) = s(b_2, B, \widehat{\lambda})$ and $s(b_1, B, \widehat{\lambda}) = s(b_2, B, \lambda)$. The details are omitted. ■

Proof of Proposition 2

We start by showing by induction that, for a generalized Borda rule, scores satisfy

$$s(a, A, \pi) = \alpha_{|A|} \sum_{a' \in A \setminus \{a\}} s(a, \{a, a'\}, \pi), \tag{14}$$

that is, the score of an alternative $a \in A$ is proportional to the sum of scores obtained by in pairwise contests with the other alternatives in A. We proceed by induction over the size of A. For $|A| = 3$ this is immediate: The Borda score a is the simple average of the score of a against the other two alternatives in A, so $\alpha_3 = \frac{1}{2}$. Our induction hypothesis is that (14) holds for any set B such that $|B| = k$. Let A be a set of size $k+1$. From the main text, we have that $s(a, A, \pi) = \frac{1}{k-1} \sum_{B \in K(a,A)} s(a, B, \pi)$ and using the induction hypothesis $s(a, B, \pi) = \alpha_k \sum_{a' \in B \setminus \{a\}} s(a, \{a, a'\}, \pi)$. Combining these we get

$$s(a, A, \pi) = \frac{\alpha_k}{k-1} \sum_{B \in K(a,A)} \sum_{a' \in B \setminus \{a\}} s(a, \{a, a'\}, \pi) = \left(\frac{k-2}{k-1}\right) \alpha_k \sum_{a' \in A \setminus \{a\}} s(a, \{a, a'\}, \pi)$$

where we used the fact that each alternative $a' \neq a$ appears in $k-2$ of the $k-1$ sets in $K(a, A)$. We conclude that (14) is satisfied for A setting $\alpha_{k+1} = \left(\frac{k-2}{k-1}\right) \alpha_k$. This completes the induction.

Note that the linearity of the scores $s(a, A, \lambda)$ in λ implies that (14) holds changing π for any population λ. Note also that $s(a, \{a, a'\}, \lambda)$ is simply the strength of motivations that prefer a to a' for population λ.

Towards a contradiction, suppose that there exists an explanation of c based on a generalized Borda count satisfying (i) and (ii) of the Proposition's hypothesis. This means that $s(a, A, \lambda) < s(x_l, A, \lambda)$ for all $a \neq x_l$ and all explanatory population λ. Now, if c satisfies (i) and (ii) in the Proposition's hypothesis, it must be that for any $\lambda \in E(c)$ we $s(x_f, \{x_f, a'\}, \lambda) \geq 1/2$ for all $a' \in A \setminus \{x_f\}$ and $s(x_l, \{x_l, a'\}, \lambda) \leq 1/2$ for all $a' \in A \setminus \{x_l\}$. It follows from (14) that

$$s(x_f, A, \lambda) \geq \frac{|A| - 1}{2} \alpha_{|A|} \geq s(x_l, A, \lambda)$$

and we have the desired contradiction. ■

B.2 Ordinal Welfare Inferences

Proof of Theorem 3

Suppose that c is pairwise coherent w.r.t. (x, y). We establish the result by constructing a family of explanations based on Plurality with populations that have support $\mathcal{S}_{xy} = \{\pi \in \Pi | x\pi y \text{ and } \nexists a \text{ such that } x\pi a\pi y\}$, i.e., the subset of Π_{xy} consisting of rankings that have no alternatives ranked between x and y. It is possible to extend the construction for almost any voting rule v, the details are omitted.

We start by introducing some notation. Let $\overline{x} \notin X$ be an auxiliary alternative that replaces x and/or y in each subset that originally contained one or both of these. To be precise, let $\overline{X} = X \backslash \{\overline{x}\}$ and consider map $g : X \to \overline{X}$

$$g(a) = \begin{cases} a & \text{if } a \notin \{x, y\} \\ \overline{x} & \text{if } a \in \{x, y\}. \end{cases}$$

For each $A \in \mathcal{A}$, $g(A)$ denotes the image of A under g. Observe that $g(A) = A$ if A does not contain either x or y and, otherwise, $g(A) = A \backslash \{x, y\} \cup \{\overline{x}\}$. Let $\overline{\mathcal{A}}$ be denote the collection of subsets of $\overline{X}$ of two or more elements, $\overline{\Pi}$ be the set of orderings over $\overline{X}$, and $\Delta^{\overline{\Pi}}$ be the simplex of populations on $\overline{\Pi}$.

Next, we introduce an auxiliary choice function c^* defined on the "reduced" domain $\overline{\mathcal{A}}$ rather than $\mathcal{A}$. For each set T that does not contain either x or y, let

$$\begin{aligned} Q(T, c) &= \{c(T \cup \{x\}), c(T \cup \{y\}), c(T \cup \{x, y\})\} \text{ and} \\ Q^g(T, c) &= g(Q(T, c)). \end{aligned}$$

Hence, $Q(T, c)$ is the set of choices under c for subsets that add x, y, or both to T and $Q^g(T, c)$ is the image of this set under g. Observe that that the image of $T \cup K$ for each nonempty $K \subseteq \{x, y\}$ under g is always $T \cup \{\overline{x}\}$.

Consider the choice function $c^* : \overline{\mathcal{A}} \to \overline{\mathcal{A}}$ defined by

$$c^*(\overline{A}) = \begin{cases} c(\overline{A}) & \text{if } \overline{x} \notin \overline{A} \\ Q^g(\overline{A} \backslash \{\overline{x}\}, c) & \text{if } \overline{x} \in \overline{A}. \end{cases} \tag{15}$$

The choice function c^* reproduces c for choice instances that do not contain $\overline{x}$ and imposes a tie for alternatives in $Q^g(\overline{A}\backslash\{\overline{x}\}, c)$ if $\overline{x} \in \overline{A}$. Note that $Q^g(\overline{A}\backslash\{\overline{x}\}, c)$ is a singleton if c is independent w.r.t. $\{x, y\}$.

For any $\lambda^* \in \Delta^{\overline{\Pi}}$ that explains c^*, we construct and an explanation of c with the desired property. For this purpose, we introduce the one-to-one map $h : \overline{\Pi} \to \mathcal{S}_{xy}$ that assigns to each preference profile $\overline{\pi} \in \overline{\Pi}$ the ranking $\pi = h(\overline{\pi}) \in \mathcal{S}_{xy}$ satisfying $a\ \overline{\pi}\ \overline{x} \Rightarrow a\pi x$ and $\overline{x}\ \overline{\pi}\ a \Rightarrow y\pi a$. That is, the order $h(\overline{\pi})$ is such that alternatives x and y are inserted as a stack in place of $\overline{x}$ respecting order $\overline{\pi}$. Let $H : \Delta^{\overline{\Pi}} \to \Delta^{\Pi}$ be the map defined by $\lambda = H(\overline{\lambda})$ if $\lambda(h(\overline{\pi})) = \lambda(\overline{\pi})$. By construction, $H(\overline{\lambda}) \in \cdot_{xy}$ for any $\overline{\lambda}$. We claim that if λ^* explains c^* then $H(\lambda^*)$ explains c. The claim is shown in two steps:
Step 1: If $\lambda = H(\overline{\lambda})$, $\overline{a} = g(a)$, and $\overline{A} = g(A)$ then

$$s(a, A, \lambda) = \begin{cases} \widehat{s}(\overline{a}, \overline{A}, \overline{\lambda}) & \text{if } \{x, y\} \subsetneq A \text{ or } a \neq y \\ 0 & \text{if } a = y,\ \{x, y\} \subseteq A, \end{cases} \tag{16}$$

where $s(a, A, \lambda)$ is the Plurality score of a from alternatives in $A \subseteq X$ and $\lambda \in \Delta^{\Pi}$, and $\widehat{s}(\overline{a}, \overline{A}, \overline{\lambda})$ is the Plurality score of $\overline{a}$ from alternatives in $\overline{A} \subseteq \overline{X}$ and $\overline{\lambda} \in \Delta^{\overline{\Pi}}$.

Indeed, we can express A as $A = T \cup K$, where $T \cap \{x, y\} = \emptyset$ and either $K = \emptyset$ or K is a non-empty subset of $\{x, y\}$. If $K = \emptyset$, so that $\overline{A} = A$, the ranking of alternatives in A under $h(\overline{\pi})$ is the same as the ranking of alternatives in $\overline{A}$ under $\overline{\pi}$ (for any $\overline{\pi}$). Instead, if $K = \{z\}$ with $z \in \{x, y\}$, the ranking of alternatives in A under $h(\overline{\pi})$ is the same than the ranking of alternatives in $\overline{A} = T \cup \{\overline{x}\}$ under $\overline{\pi}$ replacing z for $\overline{x}$ (for any $\overline{\pi}$). Thus, $s(a, A, \lambda) = \widehat{s}(\overline{a}, \overline{A}, \overline{\lambda})$ if $\{x, y\} \subsetneq A$.

Finally, consider the case $K = \{x, y\}$, so that $\overline{A} = A\backslash K \cup \{\overline{x}\}$. Since x is ranked better than y by $h(\overline{\pi})$ for any $\overline{\pi} \in \overline{\Pi}$, y is not ranked first in A by any motivation in the support of λ. Thus, for Plurality, $s(y, A, \lambda) = 0$. On the other hand, if a is ranked first in A by $h(\overline{\pi})$ then $\overline{a} = g(a)$ is ranked first in $\overline{A}$ by $\overline{\pi}$. Hence, $s(a, A, \lambda) = \widehat{s}(\overline{a}, \overline{A}, \overline{\lambda})$ for $a \neq y$.
Step 2: If A does not contain either x or y, g defined on A is the identity and, from (16), there is also a one-to-one correspondence between the scores of $H(\lambda^*)$ for alternatives in A and the scores of λ^* alternatives in $g(A)$. Further $c(A) = c^*(\overline{A})$. Since λ^* explains c^*, $\widehat{s}(c^*(\overline{A}), \overline{A}, \lambda^*) > \widehat{s}(\overline{a}, \overline{A}, \lambda^*)$ for each $\overline{a} \neq c^*(A)$, from which $s(c(A), A, H(\lambda^*)) > s(a, A, H(\lambda^*))$ for each $a \neq c(A)$. Consider the case either x or y are in A, so that $\overline{A} = g(A)$ contains $\overline{x}$. If A contains only one of these alternatives then, from (15) and (16), it is

straightforward to verify that the scores of alternatives in $A \cap Q(\overline{A}\backslash\{\overline{x}\}, c)$ are equal to each other and greater than the score of any other alternative in A. Since $c(A) \in Q(\overline{A}\backslash\{\overline{x}\}, c)$, we have that $s(c(A), A, H(\lambda^*)) \geq s(a, A, H(\lambda^*))$ for each $a \neq c(A)$. Finally, if $\{x, y\} \subseteq A$, the previous holds replacing $Q(\overline{A}\backslash\{\overline{x}\}, c)$ with $Q(\overline{A}\backslash\{\overline{x}\}, c)\backslash\{y\}$. We conclude that $H(\lambda^*) \in E(c)$. ■

Proof of Proposition 3

Suppose that c is not pairwise-coherent w.r.t. (x, y) and, without loss of generality, assume that $x = c(\{x, y\})$. This means that if there exists $\lambda \in E(c)$ that puts weight only on motivations that unanimoulsy rank these alternatives, it must be that they rank x over y ($x = c(\{x, y\})$ implies a weight of at least $1/2$ on such motivations).

Since c is not pairwise coherent w.r.t. (x, y), there exists a choice instance B such that $\{x, y\} \subset B$ and $y = c(B)$. Let v be a scoring rule such that $\lambda \in E(c, v)$. Let $\Delta s(x, y, B, \pi) = s(x, B, \pi) - s(y, B, \pi)$, the difference in scores between x and y for motivation π under voting rule v. Comparing the scores of x and y in B we have that

$$\sum_{\pi \in \Pi} \Delta s(x, y, B, \pi)\lambda(\pi) \leq 0. \tag{17}$$

Let $\Pi_{xy} = \{\pi \in \Pi | x\pi y\}$ and $\Pi_{yx} = \{\pi \in \Pi | y\pi x\}$. By assumption, we have that $\lambda_{xy} = \sum_{\pi \in \Pi_{xy}} \lambda(\pi) = 1$ and $\lambda_{yx} = \sum_{\pi \in \Pi_{yx}} \lambda(\pi) = 1 - \lambda_{xy} = 0$. We can rewrite (17) as

$$\sum_{\pi \in \Pi_{xy}} \Delta s(x, y, B, \pi)\lambda(\pi) \leq 0.$$

Since $\Delta s(x, y, B, \pi) \geq 0$ for any $\pi \in \Pi_{xy}$, the previous can only be satisfied if it is satisfied with equality and

$$\Delta s(x, y, B, \pi) = 0 \text{ for each } \pi \in \Pi_{xy} \text{ such that } \lambda(\pi) > 0.$$

Given $\lambda_{xy} = 1$, there is always some $\widetilde{\pi} \in \Pi_{xy}$ having $\lambda(\widetilde{\pi}) > 0$. Let $\rho = r(x, B, \widetilde{\pi}) \in \{1, ...|B| - 1\}$ and $\rho' = r(y, B, \widetilde{\pi}) \in \{2, ...|B|\}$ be respectively the rank of x and y from alternatives in B for motivation $\widetilde{\pi}$. Since, $\widetilde{\pi} \in \Pi_{xy}$, we have that $\rho' > \rho$ and the scoring vector $\gamma_{|B|}$ for subsets of $|B|$ alternatives must satisfy is $\gamma^{\rho}_{|B|} = \gamma^{\rho+1}_{|B|} = ... = \gamma^{\rho'}_{|B|}$. ■

Proof of Proposition 4

Let $X_n = X$ and $X_j = X\backslash\{x_{j+1}(c), ..., x_n(c)\}$ for each $j \in \{2, ..., n-1\}$. Observe that $c(X_j) = x_j(c)$. Suppose that $\lambda \in E(c)$ is an explanation such that $\Delta W(F, c, \lambda) \geq 0$ for any expansion F. This condition for each expansion $F_j = (X_j, X_{j+1})$ implies that $\widetilde{W}(x_{j+1}(c), \lambda) \geq \widetilde{W}(x_j(c), \lambda)$ for all j. This means that the partial order $\succeq_c$ is the same as the order induced by $\widetilde{W}(\cdot, \lambda)$. It follows that if $\Delta W(F, c, \lambda) = \widetilde{W}(c(B), \lambda) - \widetilde{W}(c(A), \lambda) \geq 0$ for any expansion $F = (A, B)$ then $c(B) \succeq_c c(A)$, which means that c expansion-monotonic with respect to $\succeq_c$.■

B.3 Utilitarian Bounds

Proof of Proposition 5

Let $v = (\gamma_2, ..., \gamma_n)$ be a voting system and recall that the score of an alternative a for a subset A for a population λ can be written as $s(a, A, \lambda) = \gamma^t_{|A|} P_{aA} \lambda$, where P_{aA} is the $|A| \times \Pi$ matrix described in the proof of Lemma 1. Scores are bilinear in $\gamma_{|A|}$ and λ. For a fixed choice function c, let $Q^c_{aA} = P_{c(A)A} - P_{aA}$ and observe that the polytope of explanations of c based on v is defined by mixed system of linear equalities and inequalities

$$\begin{aligned} \gamma^t_{|A|} Q^c_{aA} \lambda &\geq 0 \text{ for all } a \in A\backslash\{c(A)\},\ A \in \mathcal{A} \\ e^t \lambda &= 1 \text{ (simplex constraint).} \end{aligned} \tag{18}$$

Note also that for each $k \in \{2, ..., n\}$ the voting vector γ_k is constrained by the vector inequalities

$$\gamma_k \geq e^1_k \text{ and } \gamma_k \leq e^{k-1}_k, \tag{19}$$

which defines the $(k-2)$-dimensional cube $V^k = \{x \in \mathbb{R}^k | x^1 = 1, x^k = 0,\ 0 \leq x^i \leq 1\}$. The extreme points of this cube are precisely the basic $k-$alternative voting vectors: $e^1_k, e^2_k, ...,$ and e^{k-1}_1.

Let $Z(c, r) = \min_\lambda \{r^t \lambda | \lambda \in E(c)\}$ and $Z(c, r, v) = \min_\lambda \{r^t \lambda | \lambda \in E(c, v)\}$. Clearly, since $E(c) = \cup_{v \in V} E(c, v)$, we have that $Z(c, r) = \min_{v \in V} Z(c, r, v)$. Thus,

$$Z(c, r) = \min_{(\lambda, v)} \{r^t \lambda |\ (\lambda, v) \text{ satisfy (18) and (19)}\}$$

and the corresponding Lagrange function is

$$\begin{aligned}\mathcal{L}(\lambda, v, q, p) &= r^T\lambda + \sum_{\substack{a\in A\setminus\{c(A)\}\\ A\in\mathcal{A}}} q_{aA}\left(\gamma^t_{|A|}Q^c_{aA}\lambda\right) + p\left(e^T\lambda - 1\right)\\ &\quad + \sum_{k=2}^{n}\nu^t_{+k}(\gamma_k - e^1_k) + \sum_{k=2}^{n}\nu^t_{-k}(-\gamma_k + e^{k-1}_k)\end{aligned}$$

where the q_{aA}'s and p are the Lagrange multipliers associated to (18), and ν_{+k} and ν_{-k} are vectors of Lagrange multipliers associated to (19). For each k, basic voting vectors are the extreme points of V^k. Hence, the linearity of $\mathcal{L}$ with respect to each γ_k implies that there exists a basic voting rule that attains $Z(c, r)$. ■

Proof of Theorem 4. The result is a immediate Corollary of the following Lemma:

Lemma 4 *Fix a choice rule c and a choice-varying expansion F. Then*

$$\Delta W^{\min}(F, u, c) \geq \underline{\mu}(F, c)g(u, F, c) - (1 - \underline{\mu}(F, c))L(u, F, c)$$

and

$$\Delta W^{\max}(F, u, c) \leq \overline{\mu}(F, c)G(u, FA, c) - (1 - \overline{\mu}(F, c))l(u, F, c).$$

Proof. Observe that

$$\Delta W(F, u, c, \lambda) = \Delta W^+(F, u, c, \lambda) + \Delta W^-(F, u, c, \lambda) \tag{20}$$

where

$$\Delta W^+(F, u, c, \lambda) = \sum_{\pi\in\Pi^+}\Delta u(F, c, \pi)\lambda(\pi),$$

is the change in welfare for motivations that benefit from the expansion and

$$\Delta W^-(F, u, c, \lambda) = \sum_{\pi\in\Pi^-}\Delta u(F, c, \pi)\lambda(\pi).$$

is the change in welfare for motivations hurt by it.

Clearly,

$$\mu^+(F,c,\lambda)g(u,F,c) \leq \Delta W^+(F,u,c,\lambda) \leq \mu^+(F,c,\lambda)G(u,F,c)$$

and

$$-(1-\mu^+(F,c,\lambda))L(u,F,c) \leq \Delta W^-(F,u,c,\lambda) \leq -(1-\mu^+(A,c,\lambda))l(u,F,c).$$

Adding up these inequalities and using (20) yields

$$\mu^+(F,c,\lambda)g(u,F,c) - (1-\mu^+(F,c,\lambda))L(u,F,c) \leq \Delta W(F,u,c,\lambda)$$

and

$$\Delta W(F,u,c,\lambda) \leq \mu^+(F,c,\lambda)G(u,F,c) - (1-\mu^+(A,c,\lambda))l(u,F,c)$$

The result follows from the above two inequalities. ■

Proof of Lemma 2

We use the notation introduced in Section 3.1. Suppose that c is seemingly rational. In this case the only choice-varying expansion is F from $A = \{x^s, x^t\}$ to X, which changes the choice from $x^s = c(A)$ to $x^r = c(X)$. Note that $x^r = c(\{x^r, x^s\})$. From the choice inequality corresponding to this choice situation it follows that $\mu^+(F,c,\lambda) \geq 1/2$. This lower bound can be attained by some $\widehat{\lambda} \in E(c)$. Indeed, if $\pi = x^r x^s x^t$ and $\pi' = x^s x^r x^t$ then the population $\widehat{\lambda}$ satisfying $\widehat{\lambda}(\pi) = \widehat{\lambda}(\pi') = 1/2$ with any voting rule v explains c, and $\mu^+(A,c,\widehat{\lambda}) = 1/2$. Hence, $\underline{\mu}(F_1,c) = 1/2$ and $\underline{\rho}(F_1,c) = 1$. Since this expansion is pairwise coherent, Theorem 3 implies that $\overline{\mu}(F,c) = 1$ and $\overline{\rho}(F,c) = +\infty$.

Suppose that c displays second-place choice. There are two choice-varying expansions: $F_1 = (\{x^r, x^s\}, X)$ and $F_2 = (\{x^r, x^t\}, X)$. Both change the choice from x^r to x^s. Applying Corollary 1 we find that for any of these expansions $\underline{\mu}(F,c) = 0$ is attained for the Antiplurality polytope of explanations by the population $\lambda \in E(c, Anti)$ that puts all mass on the motivation $\pi = x^r x^s x^t$. This population explains c using Antiplurality. Thus, $\underline{\rho}(F,c) = 0$. It is straightforward to check that the same $\widehat{\lambda}$ identified above (which explains c for any v) attains $\overline{\mu}(F,c) = 1/2$, so $\overline{\rho}(F,c) = 1$.

Suppose that c displays third-place choice. There are three choice-varying expansions. Expansions F_1 and F_2 from $A_1 = \{x^r, x^s\}$ respectively $A_2 =$

$\{x^r, x^t\}$ to X, change the choice from x^r to x^t. The third expansion F_3 is from $A_3 = \{x^s, x^t\}$ to X, and the choice changes from x^s to x^t. Since $c(\{x^r, x^t\}) = x^r$ and $c(\{x^s, x^t\}) = x^s$, the strength of motivations that benefit from any expansion F is bounded above by $1/2$ for any explanation of c. Using Corollary 1 we find that $\overline{\mu}(F, c) = 1/2$ is attained by the population $\lambda(\pi) = \lambda(\pi') = 1/2$ with $\pi = x^r x^t x^s$ and $\pi' = x^s x^t x^r$, which explains c with $v = Antiplurality$ (any $v \geq 1/2$). Hence, $\overline{\rho}(F, c) = 1$ for any choice-varying expansion. The same result is used to find the lower bound $\underline{\mu}(F, c) = 1/4$ for each F, from which $\underline{\rho}(F_2, c) = 1/3$. Indeed, letting $\pi = x^r x^s x^t$, $\pi' = x^r x^t x^s$ and $\pi'' = x^s x^t x^r$, the population $\lambda(\pi) = \lambda(\pi'') = 1/4$ and $\lambda(\pi') = 1/2$ explains c with $v = Antiplurality$, and attains the bound for F_1 and F_2. Similarly, the population $\lambda(\pi) = \lambda(\pi') = 1/4$ and $\lambda(\pi'') = 1/2$ also explains c with $v = Antiplurality$, and attains the bound for F_3.

Finally, suppose that c displays cyclic choice. Assume that $c(\{x, y\}) = x, c(\{y, z\}) = y, c(\{x, z\}) = z$ and $c(X) = x$ (all other cyclic choice patterns are analogue). There are two choice-varying expansions. An expansion F_1 from $A_1 = \{y, z\}$ to X changes the choice from y to x. This expansion is pairwise coherent so, by Theorem 3, we know there exists some $\lambda \in E(c)$ such that $\mu^+(F_1, c, \lambda) = 1$. Hence $c(F_1, c) = 1$ and $\overline{\rho}(F, c) = +\infty$. Since $c(\{x, y\}) = x$, the strength of motivations that prefer x to y is at least $1/2$ for any $\lambda \in E(c)$. If $\pi = yxz$ and $\pi' = zxy$, the population λ such that $\lambda(\pi) = \lambda(\pi') = 1/2$ with Antiplurality explains c and $\mu^+(F_1, c, \lambda) = 1/2$, so the bound is attained. Hence, $\underline{\mu}(F_1, c) = 1/2$ and $\underline{\rho}(F_1, c) = 1$. The second choice-varying expansion is F_2 from $A_2 = \{x, z\}$ to X, which changes the choice from z to x. Since $c(\{x, z\}) = z$, the we have that $\mu^+(F_2, c, \lambda) \leq 1/2$ for any $\lambda \in E(c)$. Using Corollary 1 we find that $\overline{\mu}(F_2, c) = 1/2$ which is attained by the population λ such that $\lambda(\pi) = \lambda(\pi') = 1/2$ with $\pi = yxz$ and $\pi' = zxy$. This λ explains c with Antiplurality. Once again, the same result we find that $\underline{\mu}(F_2, c) = 1/4$ which is achieved by a vertex of $E(c, Anti)$.■

References

[1] Afriat, Sydney N. [1967], "The Construction of Utility Functions from Expenditure Data" *International Economic Review* 8: 67-77.

[2] Ahn, David and Haluk Ergin [2007] "Framing Contingencies," working paper, UC-Berkeley.

[3] Bernheim, B. Douglas and Antonio Rangel, "Behavioral Public Economics: Welfare and Policy Analysis with Fallible Decision-Makers," in Peter Diamond and Hannu Vartianen (eds.), Institutions and Behavioral Economics, forthcoming.

[4] Deaton, Angus and John Muellbauer [1980] Economics and Consumer Behavior, Cambridge University Press.

[5] Fostel, A., H. Scarf and M. Todd [2004], "Two New Proofs of Afriat's Theorem", *Economic Theory* 24: 211-219.

[6] Forges, Francoise and Enrico Minelli [2006] "Afriat's Theorem for General Budget Sets", CESifo Working Paper Series.

[7] Fudenberg, Drew and David Levine [2006] "A Dual Self Model of Impulse Control", *American Economic Review* 96:1449-1476.

[8] Green, Jerry R. and Daniel A. Hojman [2007] "Preference Intesity, Rationality and Welfare". In preparation.

[9] Gul, Faruk and Wolfgang Pesendorfer [2001] "Temptation and Self-Control" *Econometrica* 69: 1403–1435.

[10] Gul, Faruk and Wolfgang Pesendorfer [2005] "The Case for Mindless Economics", working paper, Princeton University.

[11] Laibson, D., A. Repetto and J. Tobacman (1998), "Self-Control and Saving for Retirement", Brookings Papers on Economic Activity, 1, 91-196.

[12] Manzini, Paola, and Marco Mariotti, [2007] "Rationalizing Boundedly Rational Choice", *American Economic Review,* forthcoming.

[13] Mas-Colell, Andreu [1978] "On Revealed Preference Analysis", *Review of Economic Studies* 45: 121-131.

[14] Mas-Colell, Andreu, Whinston and Jerry Green [1995] "Microeconomic Theory", Oxford University Press.

[15] O'Donhaue, T and M. Rabin [1999] "Doing it Now or Later," *American Economic Review* 89: 103-124.

[16] Redelmeier, D. and Eldar Shafir [1995] "Medical decision making in situations that offer multiple alternatives" *Journal of the American Medical Association* 273: 302-305.

[17] Roelfsma, Peter. H.M.P. and Daniel Read [2000] "Intransitive Intertemporal Choice," *Journal of Behavioral Decision Making* 13: 161-177.

[18] Roemer, John E. [1996] "Theories of Distributive Justice," Harvard University Press.

[19] Salant, Y. and A. Rubinstein (2007), Choice with Frames, Tel-Aviv University, working paper.

[20] Saari, Donald G. [1989] "A dictionary for voting paradoxes," *Journal of Economic Theory* 48: 443–475.

[21] Saari, Donald G. [2000] "Mathematical structure of voting paradoxes II: Positional voting," *Economic Theory* 15: 55-102.

[22] Schelling, T. C. [1984] "Choice and Consequence", Harvard University Press.

[23] Shafir, Eldar, Itamar Simonson and Amos Tversky [1993] "Reason-based Choice", *Cognition* 49: 11-36.

[24] Simonoson, Itamar [1989] "Choice based on reasons: the case of attraction and compromise effects" *Journal of Consumer Research* 16: 158-174.

[25] Strotz, R.H. [1956] "Myopia and Inconsistency in Dynamic Utility Maximization," Review of Economic Studies, 23(3), 165-180.

[26] Tversky, Amos [1969] "Intransitivity of Preference", *Psychological Review* 76: 31-48.

[27] Young, H. Peyton [1975],"Social Choice Scoring Functions", *SIAM Journal on Applied Mathematics, 28, 824-838.*