Political Science and Political Economy Working Paper

Department of Government

London School of Economics

No. 8/2009

A Model of Non-Informational Preference Change

Franz Dietrich (LSE)

Christian List (LSE)

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

A Model of Non-Informational Preference Change

Franz Dietrich and Christian List*

July 20, 2009

Abstract

According to standard rational choice theory, as commonly used in political science and economics, an agent's fundamental preferences are exogenously fixed, and any preference change over decision options is due to Bayesian information learning. Although elegant and parsimonious, such a model fails to account for preference change driven by experiences or psychological changes distinct from information learning. We develop a model of non-informational preference change. Alternatives are modelled as points in some multidimensional space, only some of whose dimensions play a role in shaping the agent's preferences. Any change in these 'motivationally salient' dimensions can change the agent's preferences. How it does so is described by a new representation theorem. Our model not only captures a wide range of frequently observed phenomena, but also generalizes some standard representations of preferences in political science and economics.

*F. Dietrich, University of Maastricht and LSE; C. List, LSE. This paper was presented at the Choice Group seminar, LSE, 1/2009, the Preference Change Workshop, LSE, 5/2009, the 3rd Workshop on Decisions, Games and Logic, HEC Lausanne, 6/2009, and the 6th Conference of the Society for Economic Design, Maastricht, 6/2009. We are grateful to the participants at these events for comments and discussion.

1 Introduction

According to standard models of rational choice, there is no such thing as genuine preference change. A rational agent has fixed preferences over fundamental alternatives or outcomes, such as fully described states of the world, and any observed changes in his or her preferences over less fundamental alternatives, such as policy options, are purely information-driven: They are due to the fact that the agent has learnt new information about which fundamental outcomes are likely to result from those options. In this way, the same fundamental preferences, together with new information, lead to revised preferences at the less fundamental level. This theoretical picture is certainly elegant and parsimonious, and although it has become increasingly common in the social sciences to criticize the assumption of 'exogenously fixed preferences' (e.g., Dryzek 1992, Green and Shapiro 1994), its explanatory power should not be underestimated (e.g., Friedman 1996).

Nonetheless, it is hard to deny that there are instances of preference change which standard rational choice theory has difficulties explaining in a natural way. Sometimes agents do undergo transformations that go beyond information learning in any ordinary sense. Imagine, to give some particularly sharp examples, a capitalist businessman who, after surviving a plane crash, decides to devote his life to charity; a workoholic who, after experiencing an illness, changes his or her priorities in life; or an ageing person whose physiological changes – quite apart from the 'wisdom of age' – affect his or her lifestyle and preferences. Or imagine someone with racist preferences who, after bonding with a new neighbour of a different racial background, gives up his or her racism. To suggest that such changes are solely the result of *information* learnt from the plane crash, from the illness, from ageing or from bonding with the neighbour seems an unsatisfactory explanation. Something more fundamental appears to be going on here. This raises the question of whether standard rational choice theory can be generalized so as to account

for the possibility of more fundamental changes of preference, while retaining, as far as possible, the theory's elegance and parsimony. Our aim in this paper is to propose such a generalization and thereby to contribute to the foundations of rational choice theory in political science and economics.

The key idea is that the alternatives over which agents have preferences – such as states of the world, policy platforms, candidates, consumer goods etc. – can be characterized along several dimensions, only some of which typically play a role in shaping the agent's preferences. We call these the *motivationally salient* dimensions. For example, a voter may form his or her preferences over policy platforms just on the basis of a conventional socio-economic left-right dimension and ignore their locations on a second, religious-secular dimension, or on a third, urban-rural dimension. Similarly, an ordinary person may form preferences over different kinds of wine just on the basis of whether those wines are red or white, sweet or dry, cheap or expensive, but be oblivious to the more subtle characteristics that the wine connoisseur appreciates. When some of these further dimensions become salient for the agent, his or her preferences can change.

A change in an agent's set of motivationally salient dimensions can be triggered by external experiences or by internal psychological or physiological changes of the agent. It is distinct from learning new information. It cannot be identified, for instance, with learning where alternatives are located on the various dimensions. In our examples, the voter may always have been abstractly aware of the existence of the religious-secular and urban-rural dimensions of policies, and yet not have been motivated by them; and the wine drinker may always have had some information about a wine's acidity and oak, and yet not have been moved by these more subtle characteristics. On our proposal, what happens when the agent's set of motivationally salient dimensions changes is simply that different dimensions attain force in shaping his or her preferences. There need not be any change in the agent's beliefs

about the locations of the alternatives on those dimensions, or about anything else.

Our claim that a change in the motivational salience of dimensions cannot generally be reduced to information learning is further reinforced by the possibility that such a change, unlike the acquisition of information, can go in two directions: New dimensions may become motivationally salient, while others may cease to be so. For example, a poor person who gradually becomes richer may be infected with what is sometimes called 'affluenza' and no longer care about the poverty-related characteristics of alternatives, while suddenly paying great attention to the luxury-related ones. All this is entirely consistent with the agent's retaining all the factual information about poverty that he or she had before. The resulting preference change is hard to model in ordinary informational terms.[1]

In our new model, alternatives are represented by points in some multi-dimensional space and an agent forms his or her preferences over the alternatives on the basis of a particular set of dimensions that have motivational salience for him or her. The agent's preferences thus depend on the locations of the alternatives on the motivationally salient dimensions, but not on their locations on others. On this picture, a change in the motivational salience of

[1] A full discussion of whether preference change of the present kind could be remodelled as information learning is beyond the scope of this paper. In ongoing work, we give detailed (and largely negative) answers to this question, by providing 'microscopic' foundations for our model. As presented here, the model is 'macroscopic', leaving open *why* a dimension gains or loses motivational salience. This is intended for reasons of generality and parsimony. Generality requires us not to single out any particular 'microscopic' foundation of salience. A dimension's salience (or lack thereof) could stem from the nature of the agent's conceptualization, imagination or perception, to give just a few examples. Similarly, parsimony is often best achieved by explaining phenomena at a particularly accessible level, not necessarily the most fundamental one. Recall, for instance, how cooperative game models describe decision making at a coalitional rather than individual level, even when reductions to non-cooperative models are possible.

some dimensions can induce a preference change.[2]

We prove a general representation theorem in this framework, showing that, under some plausible conditions, the agent's preferences are representable by an additive utility function, according to which the utility of each alternative is given by the sum of the values the agent assigns to its location on each of the motivationally salient dimensions. This additive form is consistent with many widely recognized types of preferences in political and economic contexts, such as Euclidean and other distance-based preferences in spatial voting theory and Cobb-Douglas and constant-elasticity-of-substitution preferences in the theory of consumer choice and beyond, and we give some simple illustrations. Finally, we discuss how uncertainty and lack of information can be reintroduced into our model, so as to show that our model properly generalizes a paradigmatic standard model of rational choice, by capturing the possibility of non-informational as well as informational preference change.

Our approach sheds new light on a diverse set of social-scientific phenomena ranging from fairly general phenomena such as preference change as a result of new experiences or enhanced or diminished understanding to explicitly political ones such as deliberation-induced preference change (e.g., Miller 1992, Knight and Johnson 1994, Dryzek and List 2003, List, Luskin, Fishkin and McLean 2000/2006) and Rikerian heresthetics (Riker 1986, McLean 2001), the art of political manipulation by leading voters to reconceptualize the policy space in terms of different dimensions. Important related works include Stigler and Becker's work on taste acquisition, according to which an agent's preference for listening to classical music, consuming drugs, meeting friends etc. changes over time depending on past behaviour

[2] Another way to express this idea is to say that the agent views the space of alternatives through a particular lense, focusing on the projection of the space into a particular subspace rather than on the space as a whole, so that when the subspace in focus changes, the agent's preferences change accordingly.

and experiences (e.g., Stigler and Becker 1977, Becker 1996); various authors' work on the development or loss of other-regarding preferences such as sympathy, hate, reciprocity or identification (e.g., Sen 1977, 1996, Rabin 1998, Fehr and Gaechter 1998, Sethi and Somanathan 2001, Falk and Fischbacher 2006, Dietrich 2008); dynamic inconsistency in an agent's choices, which might be involved, for example, in the development of addictions (e.g., Strotz 1955-56, Hammond 1976); the endogenous determination of preferences and tastes by environmental factors such as government policies or institutions (e.g., Polak 1976, Bowles 1998, Dietrich 2009); and the effects of issue framing on individual agency (e.g., Tversky and Kahneman 1981, Benford and Snow 2000, Gold and List 2004). Our model speaks in various ways to each of the issues raised by these contributions, representing the relevant concerns affecting the agent's preferences or tastes (such as certain perceived characteristics of alternatives, the welfare of others etc.) in terms of separate dimensions of the space of alternatives and suggesting that the agent's preference or taste change stems from a change in the salience of these dimensions. Although our model allows a large number of extensions and generalizations, we here aim to give a pedagogical introduction to the central concepts and ideas, setting aside as many technicalities as possible.

2 Basic definitions

We consider an agent's preferences over some set X of mutually exclusive alternatives, for example states of the world, policy platforms, candidates, consumer goods etc. As already indicated, we assume that the set of alternatives X is some multidimensional space, such as $\mathbb{R}^k$, with $k > 0$. Each alternative $x \in X$ can thus be written as a k-tuple $x = (x_1, x_2, ..., x_k)$, with x_j representing the j^{th} characteristic of the alternative or its location on dimension j. We write $D = \{1, 2, ..., k\}$ to denote the set of *dimensions*.

Although $\mathbb{R}^k$ is the standard example of a k-dimensional space, our model is more general. Generally, we assume that X is of the form

$$X = X_1 \times X_2 \times ... \times X_k,$$

where each X_j is a connected topological space.[3] The real line $\mathbb{R}$ or any interval on the real line are the simplest instances of such spaces, but more complex instances are also conceivable, and an individual dimension X_j could even be internally multidimensional (e.g., be itself of the form $\mathbb{R}^2$ or $\mathbb{R}^3$ etc.).[4]

A set of *motivationally salient* dimensions is a subset $S \subseteq D$. Our aim is to model how the agent's preferences depend on the set S. We consider a family of preference orders over the alternatives in X, consisting of one preference order for each possible set of motivationally salient dimensions. Interpretationally, the agent need not be – and typically is not – aware of these preference orders except the one held under his or her current set of motivationally salient dimensions. For each $S \subseteq D$, we write $\succeq_S$ to denote the agent's preference order in the event that S is the set of motivationally salient dimensions. As is conventional, $\succeq_S$ is defined as a reflexive, transitive and complete binary relation on X, and we further assume that $\succeq_S$ is continuous, thus ruling out, for example, lexicographic preferences over X.[5] We write $\succ_S$ and $\sim_S$ to denote the strict preference order and the indifference relation induced by $\succeq_S$, respectively.

[3] A topological space X_j is *connected* if it cannot be partitioned into two non-empty open sets.

[4] In a companion paper, we address (among other things) the alternative case in which some of the X_js are discrete and thus do not constitute connected topological spaces. This alternative case occurs, in particular, when some of the characteristics of the alternatives are binary.

[5] Formally, $\succeq_S$ is *continuous* if, for all $x \in X$, the sets $\{y \in X : y \succeq_S x\}$ and $\{y \in X : x \succeq_S y\}$ are both topologically closed. In our companion paper, we also discuss the non-continuous case.

In the example of an election, the underlying set of dimensions D may include a socio-economic, a religious-secular and an urban-rural dimension, and any subset $S \subseteq D$ could be motivationally salient for a given voter, depending on which dimensions play a role in shaping his or her preferences. Some voters may form their preferences solely on the basis of the socio-economic dimension of policies, while others may also take into account other dimensions. Similarly, in the case of wines, the underlying set of dimensions D may include anything ranging from the white-red, sweet-dry and cheap-expensive dimensions to those of oak and acidity etc. A wine drinker's set S of motivationally salient dimensions now contains precisely those dimensions that play a role in shaping his or her preferences. In the case of a particularly inert agent, it can even happen that the set of motivationally salient dimensions is empty. How can we make the idea of motivational salience more precise? In particular, what do we mean by saying that a dimension plays a role in shaping the agent's preferences? In the next section, we address these questions in more detail.

3 General result

We introduce two axioms on the relationship between an agent's set of motivationally salient dimensions and his or her preference order. The first captures the central idea that only the salient dimensions have any motivational force in shaping the agent's preferences: Unless two alternatives are distinct with respect to some of the motivationally salient dimensions, the agent remains indifferent between them.

Axiom 1 *'Only salient dimensions motivate.' For any two alternatives $x, y \in X$ and any set of motivationally salient dimensions $S \subseteq D$, if $x_S = y_S$, then $x \sim_S y$.*[6]

[6]For each $x \in X$, we write x_S to denote the subvector of x restricted to the dimensions

The second axiom concerns the way in which the agent's preferences respond to gains or losses of motivationally salient dimensions. For pedagogical purposes, it is useful to begin with a simple, albeit unnecessarily strong variant of the axiom, before stating the 'official' axiom. The simple variant says that the agent's preference between any two alternatives may change when an additional dimension becomes motivationally salient – or when a previously salient dimension ceases to be so – *only if* those two alternatives differ on that dimension.

Axiom 2 (simple variant) *'Only dimensions on which there is a difference motivate.' For any two alternatives $x, y \in X$, any set of motivationally salient dimensions $S \subseteq D$ and any other dimension $j \notin S$, if $x \succeq_S y \not\Leftrightarrow x \succeq_{S\cup\{j\}} y$, then $x_j \neq y_j$.*

The 'official', weaker axiom says that, if *every* gain in motivationally salient dimenions changes the preference between two given alternatives, then those two alternatives must differ on *at least one* previously non-salient dimension: The preference change must stem from *some* such difference.

Axiom 2 (official variant) *For any two alternatives $x, y \in X$ and any set of motivationally salient dimensions $S \subsetneq D$, if $x \succeq_S y \not\Leftrightarrow x \succeq_{S\cup T} y$ for* every *non-empty set $T \subseteq D\backslash S$, then $x_j \neq y_j$ for* some *$j \in D\backslash S$.*

Of course, a lot could be said about our two axioms. We can interpret them either as substantive claims about how preferences are constrained by the motivational salience of dimensions within an agent, or as formal constraints on the correct demarcation of dimensions and the correct specification of the motivationally salient ones. The first, substantive interpretation requires that both preferences and motivational salience have some

in $S \subseteq D$. When $S = \varnothing$, x_S is the empty vector, and thus $x_S = y_S$ for any $x, y \in X$ in this case.

independent psychological content – preferences obviously as representations of choice dispositions, and motivational salience as capturing some choice-relevant features of the agent's conceptualization, imagination or perception of the alternatives in X. Once this psychological content is properly specified, our two axioms become empirically testable (and we hypothesize: compelling) claims about the relationship between preferences and motivational salience. The second, formal interpretation of our axioms, by contrast, is consistent with the pure representation-theoretic spirit of classical decision theory. Here the satisfaction of the two axioms is taken to be a constraint on the correct identification of dimensions and on the ascription of motivationally salient ones to an agent. For example, if two alternatives coincide on all dimensions in a given set S and yet the agent is not indifferent between them, contrary to axiom 1, then S, on this interpretation, cannot be a correct specification of the agent's set of motivationally salient dimensions. Instead, some dimensions outside S must be motivationally salient as well. Similarly, if the addition of one or several new dimensions to the set S *always* changes the agent's preference between x and y although x and y do not differ on *any* new dimension, contrary to axiom 2, then we must have demarcated the dimensions in D incorrectly. It may be necessary, for instance, to combine one or several of the new dimensions with one or several of the existing dimensions into a single 'composite' dimension to which the preference change can be attributed. Different readers may favour different interpretations of the axioms; for the purposes of this paper, however, we need not commit ourselves to one interpretation.[7]

What is the consequence of these two axioms? Surprisingly, their joint satisfaction ensures that the agent's family of preference orders can be represented in an elegant and unified way.

[7]Recall the 'macroscopic' nature of the present model, as explained in an earlier note.

Theorem 1 *Suppose there are three or more effective dimensions in D.*[8] *Then the agent's preference orders $\succeq_S$ across all possible $S \subseteq D$ satisfy axioms 1 and 2 if and only if there exist continuous value functions $v_1 : X_1 \to \mathbb{R}$, $v_2 : X_2 \to \mathbb{R}$, ..., $v_k : X_k \to \mathbb{R}$ such that, for any two alternatives $x, y \in X$ and any set of motivationally salient dimensions $S \subseteq D$,*

$$x \succeq_S y \iff \sum_{j \in S} v_j(x_j) \geq \sum_{j \in S} v_j(y_j).$$

A proof is given in the appendix. While the proof draws on classic characterization results by Debreu (1960) and Wakker (1988), our theorem operates in a completely different framework, in so far as it describes the properties of an entire family of preference orders, and how they are constrained by the motivational salience of dimensions, rather than just a single such order, as in those classic results.[9]

Our theorem shows that, under the two axioms we have introduced, the agent's preferences are representable by an additive utility function, according to which the utility of each alternative is the sum-total of its 'value' (as assessed by the agent) on all of the motivationally salient dimensions, but not on other dimensions. Formally, for each S, the utility function is of the form $u_S : X \to \mathbb{R}$, where, for each $x \in X$,

$$u_S(x) = \sum_{j \in S} v_j(x_j).$$

[8] We call a dimension $j \in D$ *effective* if $\succeq_{\{j\}}$ is not the all-indifferent order, i.e., if the motivational salience of j alone lets the agent hold a strict preference between at least one pair of alternatives in X.

[9] The strength of our characterization is nicely illustrated by the following simple combinatorial consideration. In the absence of our axioms, a family of preference orders $(\succeq_S)_{S \subseteq D}$ (each continuous) is at best representable by 2^k possible 'value functions' (one for each $S \subseteq D$), each of which, in turn, has k arguments (corresponding to the k dimensions of X). Our result reduces this to a representation in terms of only k 'value functions', each of which takes only one argument (corresponding to a single dimension of X).

For each dimension j, the value function v_j by which the agent assesses the value of each alternative on dimension j can take a number of forms: It may depend, for example, on the distance between the alternatives and some 'ideal' location on dimension j; or it may involve a weighting of that distance so as to amplify or reduce dimension j's influence compared with others; or it may be an increasing function that assigns higher values to higher locations on dimension j.[10] In section 5 below, we illustrate how broadly applicable our framework is, by showing that several common classes of preferences studied in the social sciences are of the form described by theorem 1.

It is important to emphasize that, while our 'if and only if' result requires three or more effective dimensions, the 'if' direction also holds without this restriction: Any additive utility function as just defined satisfies our two axioms, regardless of how many or few effective dimensions there are.

It is now transparent how our model can explain preference change:

Proposition 1 *A change in the set of motivationally salient dimensions S changes the function u_S and the corresponding preference order $\succeq_S$ except in the special case in which every added or removed salient dimension j has a constant value function v_j.*

In the next section, we consider two concrete examples of preference change driven by changes in the motivational salience of dimensions.

4 Two examples

Our first example concerns a change in voter preferences. It is frequently observed by political scientists that some of the most significant changes in voting behaviour in recent decades can be attributed to a change in the

[10] The functions v_1, v_2, ..., v_n are unique up to positive affine transformations of the form $v_j \mapsto \alpha v_j + \beta_j$, with a common scalar $\alpha > 0$ for all j.

salient political dimensions. In a study of partisan realignment, Miller and Schofield (2003), for example, observe that '[p]arty voting in 1960 was still primarily driven by the economic cleavage of the New Deal. Income and class variations were strong predictors of individual voting behavior, with middle-class and professional homeowners voting Republican and working-class union members voting Democratic... By 2000, however, the New Deal party alignment no longer captured patterns of partisan voting. In the intervening 40 years, the Civil Rights and Voting Rights Acts had triggered an increasingly race-driven distinction between the parties.' Discussing the work of Carmines and Stimson (1989) and Huckfeldt and Kohfeld (1989), Miller and Schofield further point out that 'racial issues had become the dominent cleavage in American politics', driving out class-based voting, and that 'racial polarities had come to subsume a variety of other social issues as well, including abortion, womens' rights, and prayer in schools.' This change in the motivationally salient political dimensions led to a noticeable change in voter preferences: Several states that were predominantly Democratic in 1960 became Republican in 2000 and *vice-versa*.

To give a simplified illustration of the mechanism underlying such a preference change, consider the two-dimensional space shown in Figure 1, with two displayed alternatives, x and y, and a voter with an attributed ideal point. Dimension 1 might represent the conventional socio-economic left-right dimension, while dimension 2 might represent a 'social values' dimension. We assume that, for each dimension j, the voter's value function v_j on locations on that dimension is simply given by the negative value of the distance from the voter's ideal location on that dimension (we could take the square of that value if we wanted to capture Euclidean distance in the underlying space $\mathbb{R}^2$, but the present simpler definition is sufficient to make the point – we discuss more general distance-based preferences in section 5). Let us begin by considering the case in which only dimension 1 is motivationally salient for

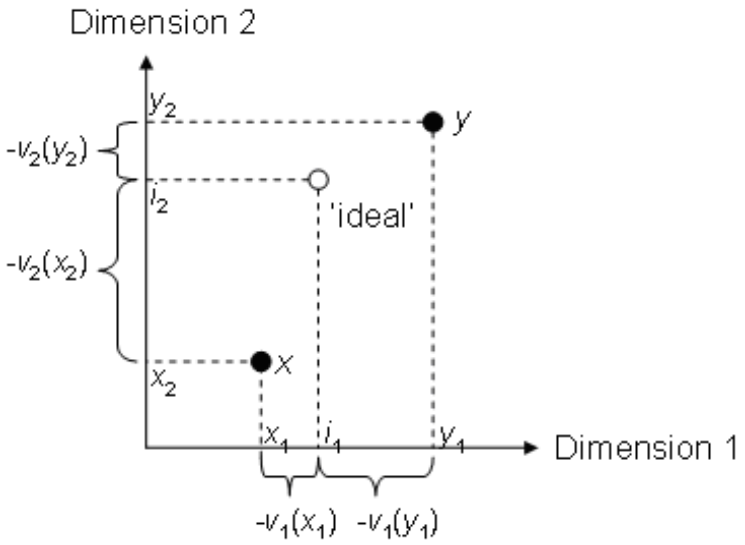

Figure 1: Voter preferences

the voter, i.e., $S = \{1\}$. Clearly, x is closer to the voter's ideal point on dimension 1 than y is, i.e.,

$$|i_1 - x_1| < |i_1 - y_1|,$$

and therefore

$$u_S(x) = v_1(x_1) = -|i_1 - x_1| > -|i_1 - y_1| = v_1(y_1) = u_S(y),$$

whence $x \succeq_S y$.

By contrast, if only dimension 2 is motivationally salient, i.e., $S = \{2\}$, we find that y is closer to the voter's ideal point on dimension 2 than x is, i.e.,

$$|i_2 - y_2| < |i_2 - x_2|,$$

which implies

$$u_S(y) = v_2(y_2) = -|i_2 - y_2| > -|i_2 - x_2| = v_2(x_2) = u_S(x),$$

and thus $y \succeq_S x$. This shows that a change in the set of motivationally salient dimensions from $S = \{1\}$ to $S = \{2\}$ can lead to a preference reversal between x and y. Moreover, in the case in which both dimensions are motivationally salient, i.e., $S = \{1, 2\}$, we also find that $y \succeq_S x$, since

$$\begin{aligned} u_S(y) &= v_1(y_1) + v_2(y_2) = (-|i_1 - y_1|) + (-|i_2 - y_2|) \\ &> (-|i_1 - x_1|) + (-|i_2 - x_2|) = v_1(x_1) + v_2(x_2) = u_S(x), \end{aligned}$$

and hence an extension of the set of motivationally salient dimensions from $S = \{1\}$ to $S = \{1, 2\}$ can also lead to the same preference reversal between x and y. Of course, real-world cases are more complex. Among other things, the positions of parties or policy alternatives may change over and above the change in motivationally salient dimensions, but the basic mechanism should be clear. As famously argued by Riker (1986), a clever political manipulator who manages to influence the perceived salience of various political dimensions can make use of such a mechanism to gain support for his or her position in the general electorate, a legislature or a committee (see also McLean 2001).

Our second example concerns preferences over consumer goods, such as cars, which may be evaluated, for instance, on the dimensions of convenience and energy efficiency. Before the issue of climate change and thereby the energy dimension became politically salient, many consumers preferred big cars, such as SUVs, to small cars, on the grounds of convenience or luxury. In the recent past, however, consumer preferences in both the United States and Europe have significantly changed, and small cars have suddenly become much more popular, while SUVs have gone out of fashion. One way to explain this preference change is by referring to a shift in the motivationally salient dimension from the convenience dimension to the energy one. This preference change cannot be reduced to information learning alone: SUVs have been known to be gas-guzzling all along, and they also remain as convenient as

ever. Of course, there is also the fact of rising energy costs, but presumably one can identify a preference change in recent years even when controlling for the cost of petrol.

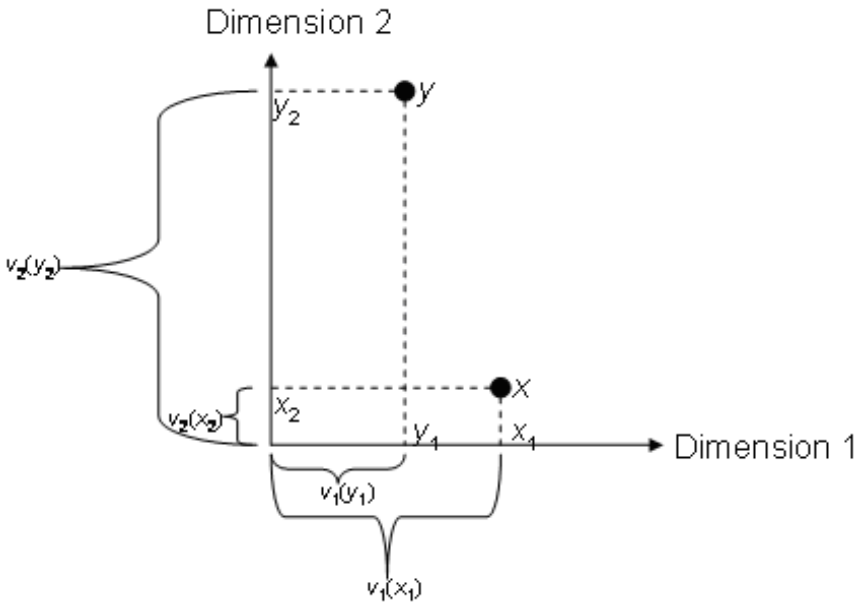

Figure 2: Consumer preferences

Figure 2 provides a stylized illustration of this scenario. The figure shows a two-dimensional space in which different cars between which a consumer can choose, such as x and y, are located. Dimension 1 might represent convenience, dimension 2 energy efficiency. Car x, the SUV, scores highly on convenience but badly on energy efficiency, and car y, the small car, has the opposite characteristics. We assume that, for each dimension j, the consumer's value function v_j on locations on that dimension is linearly increasing (for simplicity, we assume that it is given by $v_j(x_j) = x_j$, but other, more sophisticated functional forms are possible, as discussed in section 5). Again, we begin by looking at the case in which only dimension 1 is motivation-

ally salient for the consumer, i.e., $S = \{1\}$. Since x scores more highly on dimension 1 than y does, i.e., $x_1 > y_1$, we have

$$u_S(x) = v_1(x_1) = x_1 > y_1 = v_1(y_1) = u_S(y),$$

whence $x \succeq_S y$.

On the other hand, if only dimension 2 is motivationally salient, i.e., $S = \{2\}$, then y scores more highly than x does, i.e., $y_2 > x_2$, and thus

$$u_S(y) = v_2(y_2) = y_2 > x_2 = v_2(x_2) = u_S(x),$$

which implies $y \succeq_S x$. As in the earlier example, a change in the set of motivationally salient dimensions from $S = \{1\}$ to $S = \{2\}$ leads to a preference reversal between x and y. It is also easy to see that, if both dimensions are motivationally salient, i.e., $S = \{1, 2\}$, we get $y \succeq_S x$ as well, since

$$u_S(y) = v_1(y_1) + v_2(y_2) = y_1 + y_2 > x_1 + x_2 = v_1(x_1) + v_2(x_2) = u_S(x).$$

Once again, real-world cases are likely to be more complex, but our example should illustrate, in a particularly distilled form, the basic mechanism that is in operation in a broad range of cases.

5 Areas of application

So far we have only given relatively simple examples of individual preferences consistent with our model of preference change and the conditions of our representation theorem. In particular, everything has been linear in these examples. It is therefore worth going through some widely recognized, more realistic types of preferences in political and economic contexts, in order to see whether they also fit the conditions of our theorem. If they do, as we show, this underlines the wide applicability of our model to many standard social-scientific phenomena. We begin by looking at distance-based

preferences, which are familiar from spatial voting theory; we then turn to Cobb-Douglas preferences from the theory of consumer choice; and finally, we consider constant-elasticity-of-substitution (CES) preferences, also originally from consumer theory. The latter two kinds of preferences, however, are also relevant well beyond consumer theory. Cobb-Douglas preferences, for example, have been used in areas as far removed from consumer theory as international relations, to model the preferences of state actors over bundles of different goods (e.g., military goods versus civilian goods) (Oren 1994), and CES preferences have been used in the area of environmental politics, to model the idea that people have a constant elasticity of substitution between income and environmental quality, which affects their willingness to support environmental policies (Jackson 1983).

5.1 Distance-based preferences

Distance-based preferences capture the idea that an agent has a most preferred alternative, such as a most preferred policy platform or a most preferred election candidate, and prefers other alternatives less as they get more 'distant' from that most preferred preference. Accordingly, alternatives are represented by points in a multidimensional space $X = \mathbb{R}^k$ endowed with some distance metric. The space might contain all possible policy positions, and the metric could be the standard Euclidean metric or another, more general one.

Formally, a distance-based preference order is represented by a utility function $U : X \to \mathbb{R}$ according to which the utility of an alternative decreases with increasing distance from the agent's most preferred alternative or 'ideal point'. For each $x \in X$,

$$U(x) = -\left(\sum_{j=1}^{k} a_j |x_j - z_j|^p\right)^{1/p},$$

where $z = (z_1, ..., z_k)$ is the agent's most preferred alternative or 'ideal point' in X, $a_j \geq 0$ is the weight assigned to each dimension j, and $p \geq 1$ is the parameter specifying the 'degree' of the metric, often chosen to be 1 (for the 'Hamming' or 'city-block' distance) or 2 (for the 'Euclidean' distance).[11]

Suppose now, in our model, the preference order $\succeq_S$ is induced by the utility function

$$U_S(x) = -\left(\sum_{j \in S} a_j |x_j - z_j|^p\right)^{1/p},$$

which captures the natural idea that salient dimensions have positive weight, while non-salient dimensions have zero weight. Then our two axioms are clearly satisfied: The strictly increasing transformation $t \mapsto -(-t)^p$, which preserves the induced preference order $\succeq_S$, converts $U_S(x)$ into an additive utility function

$$u_S(x) = \sum_{j \in S} v_j(x_j),$$

with each value function v_j given by $v_j(x_j) = -a_j|x_j - z_j|^p$. This is precisely of the form described in theorem 1. Thus our model of preference change is applicable, for example, to standard spatial voting contexts, where it can explain such phenomena as Rikerian heresthetics or partisan realignment, as illustrated in Miller and Schofield's (2003) case study and our simple example above.

5.2 Cobb-Douglas preferences

While distance-based preferences are based on the existence of a (finite) most preferred alternative, we now turn to two classes of preferences which capture the idea that 'more is better' in each dimension, as commonly assumed in

[11] In the special case of equal weights $a_j = 1$, the utility function $U(x)$ reduces to $-\|x - z\|_p$, where $\|\bullet\|$ is the p-norm.

consumer theory and illustrated in our stylized example of consumer preferences over cars. In this subsection, we discuss Cobb-Douglas preferences, and in the next constant-elasticity-of-substitution preferences. As noted above, such preferences are also relevant well beyond consumer theory narrowly construed. The assumption that 'more is better' is plausible not only for dimensions representing consumption, but also for ones representing health, recognition by colleagues, affection by friends, wine quality etc. Even sympathetic or other-regarding preferences, as famously discussed by Sen (1977), fall into this category: Each dimension $j \in D$ could represent the welfare of one particular individual among a set of k individuals named $1, 2, ..., k$, and the 'weight' of each dimension, as formally defined in a moment, captures how much the agent represented by our model cares about that individual.

A Cobb-Douglas preference order is defined on the space of alternatives $X = (0, \infty)^k$ and represented by the utility function $U : X \to \mathbb{R}$ according to which the utility of an alternative is a weighted product of its position on each dimension. Formally, for each $x \in X$,

$$U(x) = \prod_{j=1}^{k} x_j^{a_j},$$

where $a_j \geq 0$ is the weight assigned to each dimension j.

Suppose, in our model, the preference order $\succeq_S$ is induced by the utility function

$$U_S(x) = \prod_{j \in S} x_j^{a_j},$$

again with the inbuilt stipulation than non-salient dimensions have zero weight. Then our two axioms are once again satisfied: The strictly increasing transformation $t \mapsto \log t$ converts $U_S(x)$ into an additive utility function

$$u_S(x) = \sum_{j \in S} v_j(x_j),$$

with each value function v_j given by $v_j(x_j) = a_j \ln x_j$, consistently with theorem 1. This shows that, as in our stylized example of preference change over cars, our model can explain the change of standard consumer preferences through changes in the salience of dimensions.

5.3 Constant-elasticity-of-substitution preferences

While Cobb-Douglas preferences are initially defined by a multiplicative utility function, constant-elasticity-of-substitution (CES) preferences have a more explicitly additive form. They are defined on the space of alternatives $X = [0, \infty)^k$ (the only difference to the Cobb-Douglas case being the inclusion of the zero margin) and represented by the utility function $U : X \to \mathbb{R}$ according to which the utility of an alternative is essentially the 'distance' from the zero point, as defined by an appropriate (generalized) metric.[12] For each $x \in X$,

$$U(x) = \left(\sum_{j=1}^{k} a_j x_j^p \right)^{1/p},$$

where, as before, $a_j \geq 0$ is the weight assigned to each dimension j and $p > 0$ is a parameter (interpretable as the 'degree' of the generalized metric), which is often chosen to be less than or equal to 1, so as to define a convex preference order.[13]

If, in our model, the preference order $\succeq_S$ is induced by the utility function

$$U_S(x) = \left(\sum_{j \in S} a_j x_j^p \right)^{1/p},$$

once more with the inbuilt stipulation than non-salient dimensions have zero weight, then our two axioms hold, because the strictly increasing transfor-

[12] To be precise, the generalized metric becomes a proper metric when $p \geq 1$.

[13] In the special case of equal weights $a_j = 1$, the utility function $U(x)$ reduces to $\|x\|_p$, where $\|\bullet\|$ is the p-norm.

mation $t \mapsto t^p$ converts $U_S(x)$ into a proper additive utility function

$$u_S(x) = \sum_{j \in S} v_j(x_j),$$

with each value function v_j given by $v_j(x_j) = a_j x_j^p$, as required. This completes our illustration that several standard classes of preferences familiar from the social sciences are compatible with our model of non-informational preference change.

6 Reintroducing uncertainty

At first sight, it may appear that uncertainty – the agent's lack of complete information about which fundamental alternatives or outcomes will result from his or her choices – has been banned from our model. Indeed, so far, nothing in our model is probabilistic. While the agent's preferences can change as a result of changes in his or her set of motivationally salient dimensions, probabilities do not come into play anywhere, and hence there appears to be no scope for representing uncertainty or information-driven preference change. Contrary to this appearance, we now want to show that our model has the full flexibility to represent those classical phenomena as well. In short, our model is a proper generalization of a paradigmatic standard model of rational choice, in so far as it can capture both informational and non-informational preference change.

To reintroduce uncertainty into our model, we make a theoretical move common in decision theory. We assume that the agent's preference order is defined not merely over the alternatives in the set X, but over all possible lotteries on X. Each such lottery captures one particular (subjective) probability distribution over the alternatives in X. Different lotteries could thus represent either different possible choices the agent can make, which might

have different likely consequences, or different beliefs about what the likely consequences of a single choice might be.

Let $L(X)$ denote the set of all possible lotteries, i.e., probability distributions, on X, for simplicity restricted to those with finite support,[14] and let $\tilde{\succeq}_S$ denote the agent's preference order (a reflexive, transitive and complete binary relation) over $L(X)$, for any set of motivationally salient dimensions $S \subseteq D$. As is standard, we assume that $\tilde{\succeq}_S$ is representable by the expectation of some utility function from X into $\mathbb{R}$.

The order $\tilde{\succeq}_S$ naturally induces a preference order $\succeq_S$ over the original set of alternatives X, interpreted as the set of 'sure' lotteries assigning probability 1 to a single element of X. We assume that the induced preference order $\succeq_S$ is continuous, as before.

Can we obtain a representation theorem similar to our earlier theorem, which applies to preference orders over lotteries? The following result holds.

Theorem 2 *Suppose there are three or more effective dimensions in D. Then the agent's preference orders $\tilde{\succeq}_S$ across all possible $S \subseteq D$ satisfy axioms 1 and 2 (restricted to X) if and only if there exist continuous value functions $v_1 : X_1 \to \mathbb{R}$, $v_2 : X_2 \to \mathbb{R}$, ..., $v_k : X_k \to \mathbb{R}$ such that, for any set of motivationally salient dimensions $S \subseteq D$, $\tilde{\succeq}_S$ is representable by the expectation of a utility function $\tilde{u}_S = \phi_S \circ u_S$, where*

- *$u_S : X \to \mathbb{R}$ is of the additive form $u_S(x) = \sum_{j \in S} v_j(x_j)$ for each $x \in X$,*
- *$\phi_S : u_S(X) \to \mathbb{R}$ is a strictly increasing transformation.*[15]

[14] Of course, the result to be presented could be generalized further to include the case of infinite support.

[15] The set $u_S(X)$ on which ϕ_S is defined is an interval (by the continuity of u_S and the connectedness of X), but the image of ϕ_S need not be an interval (as ϕ_S need not be continuous).

This theorem is a natural extension of our earlier theorem. While in the earlier, non-probabilistic case the agent's preference order $\succeq_S$ for each set of motivationally salient dimensions S is directly representable by an additive utility function u_S, in the probabilistic case the preference order $\tilde{\succeq}_S$ is representable by the *expectation* of a composite function $\tilde{u}_S = \phi_S \circ u_S$. This function, in turn, results from the application of a strictly increasing transformation ϕ_S to an underlying additive utility function u_S. The transformation ϕ_S can be interpreted as reflecting the agent's risk attitude for each $S \subseteq D$. Typically, ϕ_S is concave, convex, or affine (i.e., of the form $z \mapsto az + b$) depending on whether the agent is risk averse, risk loving, or risk neutral.

Note that this representation is still quite permissive. It allows not only every possible choice of continuous value functions $v_1, v_2, ..., v_k$ within the underlying additive utility function u_S,[16] but also every possible choice of strictly increasing transformations ϕ_S across $S \subseteq D$ to represent the agent's risk attitudes. In particular, the transformations ϕ_S (for $S \subseteq D$) need not even be continuous, and they can also depend on the set of motivationally salient dimensions S. Thus it is perfectly consistent with our axioms, for instance, that the agent is risk loving for some sets of motivationally salient dimensions and risk averse for others. However, if the satisfaction of axioms 1 and 2 is extended to the set of all lotteries $L(X)$ (rather than restricted to X), the transformations ϕ_S are all constrained to be positive affine and can thus be dropped, so that the representation in theorem 2 reduces to the exact counterpart of the one in theorem 1 above.

Now it is easy to see how uncertainty and information-driven preference change can be represented in our model, along with non-informational preference change, i.e., preference change from a gain or loss in motivationally salient dimensions. The agent's preference order over some choice options

[16] The existence of three or more effective dimensions merely requires that at least three of the functions $v_1, ..., v_k$ are non-constant.

changes as a result of new information whenever the agent revises the lotteries by which he or she represents those options. Ordinarily, this revision is done through Bayesian updating: The relevant probability distributions after learning the new information are obtained from the ones before learning it via Bayes's rule (or a suitable generalization of it), as in standard rational choice theory.

The key lesson of our extended representation theorem, however, is that our model retains the conceptual resources of standard rational choice theory, while also capturing the hitherto unrecognized possibility of non-informational preference change.

7 Concluding remarks

Folk psychology has long recognized the possibility of non-informational preference change. Indeed, none of the examples of non-informational preference change given in this paper should strike a non-academic reader (or indeed a successful politician who understands how political preferences can be affected by issue salience) as particularly surprising or controversial. And yet, standard rational choice theory – setting aside some of the notable exceptions cited in our introduction – adamently denies the possibility of this kind of preference change, proposing instead that every instance of preference change should be explained in purely informational terms. Even a very sophisticated review of political-science research on preference formation, which acknowledges the challenges that, for example, Tversky and Kahneman's findings on framing effects pose for classical rational choice theory, shows little willingness to give up the assumption that preference changes must always be information-driven:

> 'For the many substantive domains in which information changes do not induce preference change, no external validity is sacrificed

> by using the traditional modeling assumptions.' (Druckman and Lupia 2000, p. 13).

Although the question of how far the kind of preference change we have discussed here can be remodelled informationally deserves more comprehensive treatment (which we provide elsewhere), we wish to make two immediate remarks in response to it.

Firstly, there are some formal barriers to remodelling what we describe as non-informational preference change within the standard model of rational choice. For a start, is it difficult to capture the kind of preference change associated, for instance, with dynamic inconsistency or the loss of salient dimensions in standard terms (because their modelling would involve a violation of classical Bayesian rationality), while they do not create any special difficulties in our model. In addition, any informational remodelling of such preference change requires an enrichment of the space of alternatives or states of the world over which the agent is assumed to hold beliefs and preferences. Thus the cost of keeping the assumption of fixed underlying preferences is the inflated and often unnaturally complex ontology of alternatives or states of the world that must be ascribed to the agent.

Secondly, and perhaps even more importantly, a good theory of rational choice ought to be psychologically plausible, and this is where, in our view, the key strength of our model lies. Many frequently observed instances of apparently non-informational preference change can be explained by our model in extremely natural terms, as we hope to have illustrated. Phenomena which, from a classical vantage point, may come across as significant violations of rationality and require a major explanatory stretch, reemerge in our model as natural consequences of a change in the agent's set of motivationally salient dimensions.

Why, then, is there such a strong insistence on fixed underlying preferences in standard rational choice theory? We suspect that this simply stems

from the lack of an elegant and parsimonious model that preserves the many important and powerful insights of standard rational choice theory while also capturing the possibility that certain non-informational experiences or psychological or physiological changes may affect an agent's preferences as well. We hope that the present paper contributes to filling this gap in the literature.

Where should the present model be taken from here? Obvious extensions and generalizations of the model include the introduction of degrees of salience (as opposed to the present on-off notion of salience), the representation of discrete (e.g., binary) characteristics of alternatives (as opposed to the present focus on continuous characteristics), the consideration of weaker axioms (so as to obtain representation theorems that are more permissive than our present main theorem) and the representation of more general forms of bias, limited imagination or limited conceptualization in how an agent forms his or her preferences over a given space of alternatives. All of these are the subject of ongoing further work.

8 References

Becker, G. S. 1996. *Accounting For Tastes.* Cambridge, Mass.: Harvard University Press.

Benford, R. D. and D. A. Snow. 2000. 'Framing Processes and Social Movements: An Overview and Assessment.' *Annual Review of Sociology* 26: 611-639.

Bowles, S. 1998. 'Endogenous preferences: the cultural consequences of markets and other economic institutions.' *Journal of Economic Literature* 36 (1): 75-111.

Carmines, E. G. and J. A. Stimson. 1989. *Issue Evolution: Race and the Transformation of American Politics.* Princeton: Princeton University

Press.

Debreu, G. 1960. 'Topological methods in cardinal utility theory.' In K. J. Arrow, S. Karlin and P. Suppes, eds. 1959. *Mathematical methods in the social sciences*, pp. 16-26. Palo Alto: Stanford University Press.

Dietrich, F. 2008. 'Modelling change in individual characteristics: an axiomatic framework.' Working paper, University of Maastricht.

Dietrich, F. 2009. 'Anti-terrorism policies and the risk of provoking.' Working paper, London School of Economics.

Druckman, J. N. and A. Lupia. 2000. 'Preference Formation.' *Annual Review of Political Science* 3: 1-24.

Dryzek, J. S. 1992. 'How Far is It from Virginia and Rochester to Frankfurt? Public Choice as Critical Theory.' *British Journal of Political Science* 22 (4, Oct.): 397-417.

Dryzek, J. and C. List. 2002. 'Social Choice Theory and Deliberative Democracy: A Reconciliation.' *British Journal of Political Science* 33 (1, January): 1-28.

Falk, A. and U. Fischbacher. 2006. 'A theory of reciprocity.' *Games and Economic Behavior* 54 (2, February): 293-315.

Fehr, E. and S. Gächter. 1998. 'Reciprocity and economics: The economic implications of *Homo Reciprocans*.' *European Economic Review* 42 (3-5, May): 845-859.

Friedman, J., ed. 1996. *The Rational Choice Controversy: Economic Models of Politics Reconsidered*. New Haven: Yale University Press.

Gold, N. and C. List. 2004. 'Framing as Path-Dependence.' *Economics and Philosophy* 20 (2): 253-277.

Green, D. P. and I. Shapiro. 1994. *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. New Haven: Yale University Press.

Hammond, P. J. 1976. 'Changing Tastes and Coherent Dynamic Choice.'

Review of Economic Studies 43 (1, Feb.): 159-173.

Huckfeldt, R. R. and C. W. Kohfeld. 1989. *Race and the Decline of Class in American Politics*. Urbana: University of Illinois Press.

Jackson, J. E. 1983. 'Measuring the Demand for Environmental Quality with Survey Data.' *Journal of Politics* 45 (2, May): 335-350.

Knight, J. and J. Johnson. 1994. 'Aggregation and Deliberation: On the Possibility of Democratic Legitimacy.' *Political Theory* 22 (2, May): 277-296.

Krantz, D. H., R. D. Luce, P. Suppes and A. Tversky. 1971. *Foundations of measurement, Vol. I, Additive and polynomial representations*. New York: Academic Press.

List, C., R. C. Luskin, J. Fishkin and I. McLean. 2000/2006. 'Deliberation, Single-Peakedness, and the Possibility of Meaningful Democracy: Evidence from Deliberative Polls.' Working paper, London School of Economics.

McLean, I. 2001. *Rational Choice and British Politics: An Analysis of Rhetoric and Manipulation from Peel to Blair*. Oxford: Oxford University Press.

Miller, D. 1992. 'Deliberative Democracy and Social Choice.' *Political Studies* 40 (Special Issue): 54-67.

Miller, G. and N. Schofield. 2003. 'Activists and Partisan Realignment in the United States.' *American Political Science Review* 97 (2, April): 245-260.

Oren, I. 1994. 'The Indo-Pakistani Arms Competition: A Deductive and Statistical Analysis.' *Journal of Conflict Resolution* 38 (2, June): 185-214.

Polak, R. A. 1976. 'Interdependent Preferences.' *American Economic Review* 66 (3, June): 309-320.

Rabin, M. 1998. 'Psychology and economics.' *Journal of Economic Literature* 36 (1): 11-46.

Riker, W. H. 1986. *The Art of Political Manipulation*. New Haven: Yale University Press.

Sen, A. 1977. 'Rational fools: a critique of the behavioural foundations of economic theory.' *Philosophy and Public Affairs* 6 (4, Summer): 317-344.

Sen, A. 2006. *Identity and Violence: The Illusion of Destiny.* New York: W. W. Norton.

Sethi, R. and E. Somanathan. 2001. 'Preference Evolution and Reciprocity.' *Journal of Economic Theory* 97 (2, April): 273-297.

Stigler, G. J. and G. S. Becker. 1977. 'De gustibus non est disputandum.' *American Economic Review* 67 (2): 76-90.

Strotz, R. H. 1955-56. 'Myopia and Inconsistency in Dynamic Utility Maximization.' *Review of Economic Studies* 23 (3): 165-180.

Tversky, A. and D. Kahneman. 1981. 'The Framing of Decisions and the Psychology of Choice.' *Science* 211: 453-458

Wakker, P. 1988. 'The algebraic versus the topological approach to additive representations.' *Journal of Mathematical Psychology* 32 (4, December): 421-435.

A Proofs of the theorems

Recall that $X_1, ..., X_k$ are connected topological spaces ($k > 0$), X is their Cartesian product, and $D = \{1, ..., k\}$ is the set of dimensions.

A.1 Proof of theorem 1

Let $\succeq_S$, $S \subseteq D$, be orders (i.e., reflexive, transitive and connected binary relations) on X that are further continuous with respect to the product topology on X. We present three key steps of our proof as lemmas. (These lemmas require neither the continuity assumption, nor even that the sets X_j are connected topological spaces; they could be arbitrary sets.)

Lemma 1 *Assume axiom 2. For all sets of salient dimensions $S \subseteq D$ and*

alternatives $x, y \in X$*, if* $x_{D\setminus S} = y_{D\setminus S}$ *then* $x \succeq_S y \Leftrightarrow x \succeq_D y$*.*

Proof of lemma 1. Assume axiom 2 and let $S \subseteq D$ and $x, y \in X$ with $x_{D\setminus S} = y_{D\setminus S}$. As the equivalence $x \succeq_S y \Leftrightarrow x \succeq_D y$ holds trivially if $S = D$, let us assume that $S \subsetneq D$. As D is finite, a finitely repeated application of axiom 2 yields a nested sequence of sets $S = S_1 \subsetneq S_2 \subsetneq ... \subsetneq S_m = D$ $(2 \leq m < \infty)$ such that $x \succeq_{S_k} y \Leftrightarrow x \succeq_{S_{k+1}} y$ for each $k \in \{1, ..., m-1\}$. Hence $x \succeq_S y \Leftrightarrow x \succeq_D y$. ■

For any set $S \subseteq D$, any $s \in \times_{j\in S} X_j$ and any $t \in \times_{j\in D\setminus S} X_j$, we write (s, t) to denote the vector $x \in X$ given by $x_S = s$ and $x_{D\setminus S} = t$, i.e., the vector that coincides with s and t on the dimensions in S and $D\setminus S$, respectively.

Lemma 2 *Assume axioms 1 and 2. Then* $x \succeq_S y \Leftrightarrow (x_S, z_{D\setminus S}) \succeq_D (y_S, z_{D\setminus S})$ *for all alternatives* $x, y, z \in X$ *and sets of salient dimensions* $S \subseteq D$*.*

Proof of lemma 2. Assume axioms 1 and 2 and let $x, y, z \in X$ and $S \subseteq D$. By axiom 1,

$$x \sim_S (x_S, z_{D\setminus S}) \text{ and } y \sim_S (y_S, z_{D\setminus S}). \ (*)$$

By lemma 1, $(x_S, z_{D\setminus S}) \succeq_D (y_S, z_{D\setminus S})$ is equivalent to $(x_S, z_{D\setminus S}) \succeq_S (y_S, z_{D\setminus S})$, which by (*) and the transitivity of $\succeq_S$ is equivalent to $x \succeq_S y$. ■

As usual, we call an order $\succeq$ on X *separable* if, for every set $S \subseteq D$, all $s, s' \in \times_{j\in S} X_j$ and all $t, t' \in \times_{j\in D\setminus S} X_j$, we have $(s, t) \succeq (s', t) \Leftrightarrow (s, t') \succeq (s', t')$ (i.e., if for all $S \subseteq D$ the way in which the subspace $\times_{j\in S} X_j$ is ordered given that we fix the coordinates on the dimensions in $D\setminus S$ does not depend on how these coordinates are fixed).

Lemma 3 *Assume axioms 1 and 2. The full-salience order* $\succeq_D$ *is separable.*

Proof of lemma 3. Assume axioms 1 and 2. Consider any $S \subseteq D$, $s, s' \in \times_{j\in S} X_j$ and $t, t' \in \times_{j\in D\setminus S} X_j$. We have to show that $(s,t) \succeq_D (s',t) \Leftrightarrow (s,t') \succeq_D (s',t')$. This equivalence holds because, choosing arbitrary $x, y \in X$ with $x_S = s$ and $y_S = s'$, each side of it is equivalent to $x \succeq_S y$ by lemma 2. ■

Assume now that at least three dimensions in D are effective. Firstly, the reader may easily check that if the preference orders $\succeq_S$, $S \subseteq D$, have the specified additive form, then they obey both axioms 1 and 2 (and would do so even without our assumptions of continuity and the effectiveness of at least three dimensions).

Conversely, assume axioms 1 and 2. Each effective dimension $j \in D$ is essential under the full-salience order $\succeq_D$, i.e., there is a non-indifference $x \not\sim_D y$ for at least one pair $x, y \in X$ with $x_{D\setminus\{j\}} = y_{D\setminus\{j\}}$. To see why this is the case, recall that j's effectiveness implies existence of a pair $x, y \in X$ with $x \not\sim_{\{j\}} y$, which by lemma 2 entails $x' \not\sim_D y'$ where $x' := (x_{\{j\}}, z_{D\setminus\{j\}})$ and $y' := (y_{\{j\}}, z_{D\setminus\{j\}})$ for an arbitrarily chosen $z \in X$. So, since at least three dimensions j are effective, at least three dimensions are essential under $\succeq_D$. Moreover, $\succeq_D$ is separable by lemma 3 and continuous by assumption. So, by Wakker's (1988) strengthened version of Debreu's (1960) additive representation theorem,[17] there exist continuous functions $v_j : X_j \to \mathbb{R}$, $j \in D$, such that

$$x \succeq_D y \Leftrightarrow \sum_{j\in D} v_j(x_j) \geq \sum_{j\in D} v_j(y_j) \text{ for all } x, y \in X. \tag{1}$$

[17] Debreu's (1960) original theorem uses an additional assumption (each X_j has a countable topologically dense subset), which is removed by Wakker (1988) (and earlier by Krantz et al. 1971, yet without proving the continuity of the functions in the representation).

We now need to show that, for all $S \subseteq D$ and $x, y \in X$, $x \succeq_S y$ is equivalent to

$$\sum_{j \in S} v_j(x_j) \geq \sum_{j \in S} v_j(y_j). \tag{2}$$

Consider any $S \subseteq D$ and $x, y \in X$. Fix an arbitrary $z \in X$. By lemma 2, $x \succeq_S y$ is equivalent to $(x_S, z_{D \setminus S}) \succeq_D (y_S, z_{D \setminus S})$, which by (1) is equivalent to

$$\sum_{j \in S} v_j(x_j) + \sum_{j \in D \setminus S} v_j(z_j) \geq \sum_{j \in S} v_j(y_j) + \sum_{j \in D \setminus S} v_j(z_j),$$

and hence (by cancelling out) to (2), as required. ■

A.2 Proof of theorem 2

Our proof of theorem 2 draws on theorem 1. Recall that $\tilde{\succeq}_S$, $S \subseteq D$, are orders on the set $L(X)$ of lotteries (i.e., probability distributions with finite support) over X. Each $\tilde{\succeq}_S$ is assumed to be representable by the expectation of some function $X \to \mathbb{R}$. Identifying alternatives in X with sure lotteries, we denote by $\succeq_S$, $S \subseteq D$, the induced orders on X, which are continuous by assumption.

Suppose at least three dimensions are effective. Firstly, assume axioms 1 and 2. Then, by the 'only if' part of theorem 1, there exist continuous functions $v_j : X_j \to \mathbb{R}$, $j \in D$, such that the restriction $\succeq_S$ of any $\tilde{\succeq}_S$, $S \subseteq D$, to the set X of sure lotteries is representable by the function $u_S : X \to \mathbb{R}$ given by $u_S(x) = \sum_{j \in S} v_j(x_j)$. Let $S \subseteq D$. By assumption, there exists a function $\tilde{u}_S : X \to \mathbb{R}$ whose expectation represents $\tilde{\succeq}_S$. In particular, $\tilde{u}_S$ represents the restriction $\succeq_S$ of $\tilde{\succeq}_S$ to X, the set of sure lotteries. So, $\tilde{u}_S$ represents the same order $\succeq_S$ as u_S. Hence, $\tilde{u}_S = \phi_S \circ u_S$ for some strictly increasing function $\phi_S : u_S(X) \to \mathbb{R}$, as required.

Conversely, assume that the orders $\tilde{\succeq}_S$, $S \subseteq D$, are representable in the specified way, and let $\tilde{u}_S$, ϕ_S, u_S, $S \subseteq D$, be the functions that feature in

one such representation. In particular, the restriction $\succeq_S$ of any $\tilde{\succeq}_S$ to the set X of sure lotteries is representable by $\tilde{u}_S$, hence also by u_S (as u_S and $\tilde{u}_S$ are strictly increasing transformations of each other). So, by the 'if' part of theorem 1, axioms 1 and 2 are satisfied. ■