Where do preferences come from?

Franz Dietrich and Christian List*
London School of Economics

December 2010

Abstract

Rational choice theory analyzes how an agent can rationally act, given his or her preferences, but says little about where those preferences come from. Instead, preferences are usually assumed to be fixed and exogenously given. We introduce a framework for conceptualizing preference formation and preference change. In our model, an agent's preferences are based on certain 'motivationally salient' properties of the alternatives over which the preferences are held. Preferences may change as new properties of the alternatives become salient or previously salient ones cease to be so. We suggest that our approach captures endogenous preferences in various contexts, and helps to illuminate the distinction between formal and substantive concepts of rationality, as well as the role of perception in rational choice.

Keywords: preference formation, preference change, properties, motivations, reasons, endogenous preferences, formal versus substantive rationality, perception

1 Introduction

Rational choice theory offers a powerful framework for analyzing how agents can rationally make decisions in various situations, whether alone or in interaction with others, whether under uncertainty or under complete information, whether on the basis of self-interest or on the basis of other-regarding motivations, and so on. Its key idea goes back at least to David Hume's account of rational agency from the 18th century (e.g., Hume 1739): an agent has beliefs and preferences, now usually modelled as subjective probabilities and utilities, and acts so as to satisfy his or her preferences in accordance with his or her beliefs. This model of rational agency serves as the foundation of much of economic theory and the social sciences, ranging from decision theory, game theory and social choice theory on the theoretical side to the theories of consumer choice and the firm, general equilibrium theory, spatial voting theory and even the 'realist' theory of international relations on the more applied side. Like any influential paradigm, rational choice theory has come under criticism from several angles – both theoretical and empirical – and may ultimately need to be revised or generalized, but for now it remains an indispensable part of any economist's toolbox.

However, while rational choice theory is able to show with great precision how an agent can rationally act, *given his or her preferences*, one of the standard theory's

*We are grateful for helpful discussions with David Austen-Smith, Richard Bradley, John Broome, Tim Feddersen, Philip Pettit, Wlodek Rabinowicz, and Alvaro Sandroni.

shortcomings is that it has very little to say about where those preferences come from or how they might change. An agent's preferences are assumed to be *fixed* and *exogenously given* (a few exceptional, revisionary works are cited below). They are *fixed* in that the agent never changes his or her 'fundamental' preferences over fully specified outcomes; at most, the agent may change his or her 'derived' preferences over actions after learning new information about their likely outcomes. And preferences are *exogenously given* in that the theory cannot explain how they are formed: an agent's preferences are simply taken to be an essential but inexplicable feature of the agent's personal identity.

Our aim in this paper is to introduce a formal framework for conceptualizing and modelling preference formation and preference change, and thereby to contribute to filling rational choice theory's gap in this respect. As explained in detail below, the central idea is that an agent's preferences are based on certain 'motivationally salient' properties of the alternatives over which the preferences are held. Accordingly, an agent's preferences may change as new properties of the alternatives become salient or previously salient ones cease to be so. Our paper is devoted to exploring the consequences of this idea. The present work, in turn, is part of a larger project on the role of reasons in rational choice.[1]

Of course, there are several earlier works on endogenously determined preferences,[2] and what we are proposing should not be seen as a rival to established work in rational choice theory, but as a friendly amendment, albeit an ambitious one. Crucially, the paper's contribution is conceptual and foundational, not technical, and we keep the formalism to a minimum. But despite being formally elementary, we hope to provide an interesting, and at least somewhat provocative, basis for further work.

The paper is structured as follows. In Section 2, we introduce our main framework for capturing the relationship between properties and preferences, and define what it means for an agent's preferences to be 'property-based'. In Section 3, we present an axiomatic characterization of property-based preferences, which shows that our key concept is not *ad hoc*, but justifiable in terms of two simple axioms. In Section 4, we discuss what all this suggests for preference formation and preference change. We also respond to the objection that our theory has so many degrees of freedom that it runs the risk of becoming unfalsifiable. In Section 5, we distinguish between two concepts of rationality – a thin, formal one and a thicker, substantive one – and show that our framework can be used to formalize both, thereby allowing us to clear up some common misunderstandings between economists and philosophers. In Section 6, we offer an alternative perspective on our theory, suggesting that our property-based account of preference formation can also be re-expressed as a 'double-ontology' account, in which the agent's ontology of alternatives is distinguished from that of the modeller. In Section 7, finally, we sketch a simple game-theoretic application, illustrating how our theory might be used to capture endogenous preferences in concrete cases.

[1] For a companion paper addressed to a philosophical audience, with further references to the related literature, see Dietrich and List (2009). For some related informal ideas, see Pettit (1991).

[2] For works on how human tastes and other characteristics depend on environmental factors such as institutions, policies, and interactions with other people, see, for instance, Polak (1976), Bowles (1998), Rabin (1998) and Dietrich (2008a). For explicit analyses of preference change, also in the context of dynamic inconsistency, see Strotz (1955-56), Hammond (1976), O'Donoghue and Rabin (1999), Bradley (2007), Dietrich (2008b) and Dietrich and List (forthcoming). Preference evolution is analyzed, e.g., by Dekel et al. (2007). For discussions of preference change in group deliberation, see Dryzek and List (2003) and List, Fishkin, Luskin and McLean (2000/2006).

2 The main framework

2.1 The objects of preference

We want to model how an agent forms and possibly revises his or her preferences over some set X of fundamental objects of preference. Depending on the application, the objects in X can be fully described outcomes or consequences of actions, possible worlds, social states, bundles of goods, or policy platforms. For simplicity, we call them *alternatives*. Our only assumption is that the alternatives in X are mutually exclusive and jointly exhaustive of the relevant space of possibilities. Although it is sometimes useful to define an agent's preferences not just over fundamental alternatives, but also over prospects involving uncertainty (e.g., probability distributions over alternatives), we here focus on preferences in the 'pure' case without uncertainty.

2.2 The agent's preferences

We represent the agent's preferences by some order $\succsim$ on X (a complete and transitive binary relation), where $x \succsim y$ means that the agent weakly prefers x to y. As usual, $\succ$ and $\sim$ denote the strict and indifference parts of $\succsim$. Unlike in standard rational choice theory, we do not treat the agent's preference order $\succsim$ as given, but are interested in how it is formed and how it may be revised. To address these questions, we introduce the idea that the agent's preferences over the alternatives in X depend on certain properties of those alternatives. To explain this idea, we first define properties of alternatives in general; we then define the agent's motivational state as a set of 'motivationally salient' properties of the alternatives; and finally we specify how the agent's preference order depends on those motivationally salient properties.

2.3 Properties of the alternatives

Informally, a property is a characteristic that an alternative may or may not have. For example, being fat-free or being vegetarian is a property that a dinner option may or may not have. Any property thus partitions the set X of alternatives into those that have the property and those that do not. (There are number of ways in which our approach can handle non-binary properties, but we set this issue aside for present purposes; see, e.g., Dietrich and List forthcoming.)

Formally, we can define a property either *extensionally* or *intensionally*. On the extensional approach, a property is simply defined as the subset of X containing all the alternatives that have the property. So the property of redness is nothing but the set of all red objects in X. On the intensional approach, by contrast, a property is defined in terms of some description or label, such as 'being red'. Thus a property merely *picks out* a subset of X but is not *identified* with it. This can accommodate the fact that two or more differently described properties can have the same extension. In a given set of election candidates, for instance, the properties of 'being at least 50 years old' and 'having at least 25 years of political experience' may be satisfied by the same people. It is well known at least since Kahneman and Tversky's classic works on framing effects that the description of something may affect an agent's attitude towards it (e.g., Tversky and Kahneman 1981; see also Gold and List 2004).

2.4 The agent's motivational state

The key idea underlying our approach is that, in forming his or her preferences, an agent focuses on some, but not necessarily all, properties of the alternatives. We call the properties that the agent focuses on the *motivationally salient* ones, and we call the set of such properties, M, the agent's *motivational state*. Which properties are in M in any situation is ultimately a psychological question, which a formal theory by itself cannot settle. For example, after having suffered from gallstones, an agent may form his or her preferences over various kinds of food on the basis of whether they are healthy, and may not be interested in whether their taste is rich. Or, in forming preferences over cars, an agent may focus on whether a car is cheap, safe and environmentally friendly, and may not be interested in whether it is good for racing or whether it will impress the neighbours. Someone else, on the other hand, may be preoccupied with the latter properties.

We write $\mathcal{M}$ to denote the set of all motivational states that are deemed psychologically possible for the agent; formally, $\mathcal{M}$ is some non-empty set of sets of properties. By stating which specifications of M are included in $\mathcal{M}$, we can capture different assumptions about which properties can simultaneously become motivationally salient for the agent. Such assumptions range from certain minimal richness assumptions, according to which some properties are always motivationally salient, to certain 'crowding out' or 'crowding in' assumptions, whereby the motivational salience of some properties either excludes, or necessitates, the motivational salience of others. For instance, when an agent gives too much attention to the monetary properties of something, such as the financial rewards from taking an action, he or she may lose sight of its charitable properties, such as the fact that it benefits others, as famously suggested by Titmuss's comparative study of voluntary blood donations in the UK and paid blood donations in the US (Titmuss 1970).

2.5 Property-based preferences

To indicate that the agent's preference order $\succsim$ depends on his or her motivational state M, we append the subscript M to the symbol $\succsim$, interpreting $\succsim_M$ as the agent's preference order in state M. A full model of an agent thus requires the ascription of an entire family of preference orders $\succsim_M$ to the agent, one for each motivational state $M \in \mathcal{M}$. Below we suggest a dispositional interpretation of this family.

So how exactly does $\succsim_M$ depend on M? We call the agent's family of preference orders $(\succsim_M)_{M \in \mathcal{M}}$ *property-based* if there exists a binary relation $\geq$ over property combinations (consistent sets of properties[3]) such that, for any motivational state $M \in \mathcal{M}$ and any alternatives $x, y \in X$,

$$x \succsim_M y \Leftrightarrow \{P \in M : x \text{ satisfies } P\} \geq \{P \in M : y \text{ satisfies } P\}.$$

We then say that x's having the properties in $\{P \in M : x \text{ satisfies } P\}$ and y's having the properties in $\{P \in M : y \text{ satisfies } P\}$ are the agent's *motivating reasons* (or *motives*) for preferring x to y in state M. And we call $\geq$ the agent's *weighing relation* over property combinations.

[3] A set of properties is *consistent* if there exists an alternative $x \in X$ which satisfies all of them.

Notice that, although the agent's preference order $\succsim_M$ in any motivational state $M \in \mathcal{M}$ is complete and transitive, the underlying weighing relation need not be. However, the following holds:

Remark 1 *A binary relation $\geq$ over property combinations induces a property-based family of preference orders $(\succsim_M)_{M \in \mathcal{M}}$ if and only if, for any $M \in \mathcal{M}$, the restriction of $\geq$ to the set $X_M = \{\{P \in M : x \text{ satisfies } P\} : x \in X\}$ is complete and transitive.*

The right-hand side of this biconditional is consistent with the relation $\geq$ itself not being complete or transitive; for an example, see Dietrich and List (2009). We should note that, although we here focus on preference relations that are complete and transitive, nothing in our approach rules out the analysis of incomplete or intransitive preferences. Indeed, if preferences are property-based, then their formal structure will depend on the structure of the agent's weighing relation, and if certain property combinations in the same set X_M turn out to be mutually incomparable or not transitively ranked, then this will show up in the agent's preferences.

2.6 An example

A simple example helps to illustrate the ideas just introduced. Consider an agent faced with a choice between four alternatives, namely different cakes:

S&H: a sweet and healthy cake, ¬S&H: a non-sweet and healthy cake,
S&¬H: a sweet and unhealthy cake, ¬S&¬H: a non-sweet and unhealthy cake.

For simplicity, suppose the only properties that may become motivationally salient for the agent are:

S: The cake is sweet. H: The cake is healthy.

Suppose further that any set of properties can in principle be motivationally salient, so that the set of all possible motivational states is

$$\mathcal{M} = \{\{\text{S,H}\}, \{\text{S}\}, \{\text{H}\}, \varnothing\}.$$

Now the agent's preferences across different $M \in \mathcal{M}$ might be as follows:

In state $M = \{\text{S,H}\}$: S&H $\succ_M$ nS&H $\succ_M$ S&nH $\succ_M$ nS&nH.
In state $M = \{\text{S}\}$: S&H $\sim_M$ S&nH $\succ_M$ nS&H $\sim_M$ nS&nH.
In state $M = \{\text{H}\}$: S&H $\sim_M$ nS&H $\succ_M$ S&nH $\sim_M$ nS&nH.
In state $M = \varnothing$: S&H $\sim_M$ nS&H $\sim_M$ S&nH $\sim_M$ nS&nH.

These preferences can be verified to be property-based, with respect to the following weighing relation:

$$\{\text{S,H}\} > \{\text{H}\} > \{\text{S}\} > \varnothing.$$

In short, a single weighing relation over property combinations suffices to induce the agent's entire family of preference orders across different motivational states. Needless to say, different weighing relations may lead to different preferences.

3 An axiomatic characterization

To show that our definition of property-based preferences is not *ad hoc*, it is useful to characterize such preferences axiomatically. The first of the two theorems to be presented is a variant of an earlier result in a terminologically distinct but formally equivalent setting (Dietrich and List 2009), while the second is new; proofs are given in the Appendix. The following two axioms seem to be reasonable constraints on the relationship between motivationally salient properties and preferences.

Axiom 1 *'Only motivationally salient properties motivate.' For any two alternatives* $x, y \in X$ *and any motivational state* $M \in \mathcal{M}$*,*

$$\textit{if } \{P \in M : x \textit{ satisfies } P\} = \{P \in M : y \textit{ satisfies } P\}\textit{, then } x \sim_M y.$$

This axiom simply says that the agent is indifferent between any two alternatives whose motivationally salient properties are the same.

Axiom 2 *'Adding motivationally salient properties not satisfied by either of two alternatives does not change the preference between them.' For any two alternatives* $x, y \in X$ *and any two motivational states* $M, M' \in \mathcal{M}$ *with* $M' \supseteq M$*,*

$$\textit{if neither } x \textit{ nor } y \textit{ satisfies any } P \in M' \backslash M\textit{, then } x \succsim_M y \Leftrightarrow x \succsim_{M'} y.$$

This axiom says that if the agent's motivational state is extended, in that additional properties become motivationally salient, this does not change the agent's preference between any alternatives that satisfy none of the newly added properties. This is weaker than the requirement that the preference between any two alternatives should never change so long as the newly added properties do not discriminate between them.

Axiom 3 *'Adding motivationally salient properties that do not discriminate between two alternatives does not change the preference between them.' For any two alternatives* $x, y \in X$ *and any two motivational states* $M, M' \in M$ *with* $M' \supseteq M$*,*

$$\textit{if } [x \textit{ satisfies } P \;\Leftrightarrow\; y \textit{ satisfies } P] \textit{ for any } P \in M' \backslash M\textit{, then } x \succsim_M y \Leftrightarrow x \succsim_{M'} y.$$

While this stronger requirement may be plausible if different motivationally salient properties have a separable effect on the agent's preferences, it rules out the possibility that the motivational effect of some properties may depend on which other properties are also motivationally salient. Since we do not wish to exclude such non-separability of different properties *a priori*, we generally defend only the weaker requirement, not the stronger one.

What is the consequence of our axioms? It turns out that, if the set of possible motivational states satisfies a suitable closure condition, Axioms 1 and 2 characterize the class of property-based families of preference orders. Call $\mathcal{M}$ *intersection-closed* if, whenever $M_1, M_2 \in \mathcal{M}$, then $M_1 \cap M_2 \in \mathcal{M}$.

Theorem 1 *Suppose* $\mathcal{M}$ *is intersection-closed. Then the agent's family of preference orders* $(\succsim_M)_{M \in \mathcal{M}}$ *satisfies Axioms 1 and 2 if and only if it is property-based.*

Thus the two axioms guarantee that the agent's preferences across variations in his or her motivational state are representable by a single underlying weighing relation over property combinations. If we replace Axiom 2 by Axiom 3 (and strengthen the closure condition on $\mathcal{M}$), we obtain the following stronger characterization. Call $\mathcal{M}$ *subset-closed* if, whenever $M_1 \in \mathcal{M}$ and $M_2 \subseteq M_1$ then $M_2 \in \mathcal{M}$.

Theorem 2 *Suppose $\mathcal{M}$ is subset-closed. Then the agent's family of preference orders $(\succsim_M)_{M \in \mathcal{M}}$ satisfies Axioms 1 and 3 if and only if it is property-based in a separable way, i.e., the family $(\succsim_M)_{M \in \mathcal{M}}$ is representable by a separable weighing relation.*[4]

Separability of the weighing relation means that the ranking of any property combination S_1 relative to any other S_2 does not depend on which further properties are present, *ceteris paribus*; formally,

$$S_1 \geq S_2 \text{ if and only if } S_1 \cup T \geq S_2 \cup T,$$

where T is any set of properties not in S_1 or in S_2 but consistent with each of S_1 and S_2. In general, there is no such restriction on the weighing relation.

4 What this suggests for preference formation and preference change

4.1 The basic implication

If our theory is correct, the underlying stable feature characterizing an agent is not the agent's preference order over the alternatives in X, as in standard rational choice theory, but the agent's weighing relation over property combinations. This weighing relation is an abstract entity, which need not be directly cognitively accessible to the agent. It could be interpreted, for instance, in dispositional terms. If $S_1 \geq S_2$, this could be taken to mean that the agent is disposed to prefer an alternative whose motivationally salient properties are those in S_1 to one whose motivationally salient properties are those in S_2.

While the agent's weighing relation is stable on this picture, his or her motivational state is variable. This suggests the following picture of preference formation and preference change:

- An agent *forms* his or her preferences by adopting a particular motivational state, i.e., by focusing – consciously or otherwise – on certain properties of the alternatives as the motivationally salient ones (and by taking on a weighing relation in the first place).
- An agent may *change* his or her preferences when the motivational state changes, i.e., when new properties of the alternatives become motivationally salient or others cease to be salient.

[4] The assumption that $\mathcal{M}$ is subset-closed cannot be weakened to Theorem 1's assumption of intersection-closure. A counterexample is given at the end of the Appendix.

4.2 Is this theory empirically testable?

At first sight, one might worry that the greater degrees of freedom in our model, compared to standard models of rational choice, imply that it can 'explain' almost anything, i.e., that it might be unfalsifiable. We now want to show that this is not the case and that the model can be empirically operationalized.

To do so, we need to introduce one further idea, namely that empirically observable contexts induce particular motivational states. Let us define a *context* as a situation the agent can observably be in, and let us write $\mathcal{C}$ to denote a set of contexts. A *context* might be:

- a concrete choice situation, as given by a feasible set of alternatives;
- a particular way in which a decision problem is framed (in Kahneman and Tversky's sense of framing);
- a socially well-defined role in which the agent is expected to act in a given situation;
- an observable life circumstance of the agent; and so on.

What matters is that the different contexts in $\mathcal{C}$ are empirically distinguishable. For a full operationalization of our theory, we need to add to our formal model:

(a) a hypothesis about what the agent's set of psychologically possible motivational states $\mathcal{M}$ is, as already mentioned;
(b) a hypothesis about what the relation between empirically observable contexts and motivational states is, as captured by some *motivation function* $f : \mathcal{C} \to \mathcal{M}$;
(c) a hypothesis about the agent's weighing relation $\geq$.

It should be evident that the conjunction of our model and (a), (b) and (c) straightforwardly entails what the agent's preferences will be in any given context – namely $\succsim_{f(C)}$ for context $C \in \mathcal{C}$ – and consequently what his or her choice behaviour will look like, assuming the usual relationship between preferences and choices. Hence the resulting theory is straightforwardly testable. When presented with recalcitrant evidence, of course, we will always face a choice between giving up our core model itself and giving up one or more of the 'auxiliary assumptions' under (a), (b) and (c). This predicament, however, is no different from the familiar one in other areas of science.

4.3 Minimal constraints on testability

Although a full operationalization of our theory requires a full specification of (a), (b) and (c), it is worth observing that a suitable constraint under any one of (a), (b) or (c) alone is already sufficient to render the resulting theory empirically non-vacuous. We give the simplest 'toy' examples by which it is possible to establish this point.

4.3.1 Constraining the agent's motivational states under (a)

Suppose, for instance, we hypothesize that in any psychologically possible motivational state $M \in \mathcal{M}$ there are at most three motivationally salient properties, while we do not make any hypotheses under (b) and (c). Given our model, the present hypothesis alone implies that, in any motivational state, the agent's preference order will have no more than 8 ($= 2^3$) indifference classes, which, in turn, is a testable implication, assuming, as before, the usual relationship between preferences and choices.

4.3.2 Constraining the agent's motivation function under (b)

Suppose we only hypothesize that the agent's motivation function is constant, i.e., that any context triggers the same motivational state, while saying nothing about (a) and (c). In this case, our model reduces to a variant of a standard model of rational choice, according to which the agent's preference order is context-independently fixed. Since this standard model is testable, the present model will be testable too.

4.3.3 Constraining the agent's weighing relation under (c)

Suppose, to give a rather trivial example, we hypothesize that the agent's weighing relation is dichotomous, i.e., distinguishes only between two equivalence classes of property combinations. Regardless of our assumptions about (a) and (b), this constrains the agent's preference order in any motivational state to have no more than two indifference classes, a testable implication. For a less trivial example, consider the hypothesis that the agent's weighing relation is separable. If we can find two pairs of alternatives x, y and x', y' such that x' and y' are obtained from x and y, respectively, by adding the same properties, then our model implies that, in any motivational state, the agent prefers x to y if and only if he or she prefers x' to y', a testable implication so long as the agent's motivational state is assumed to be stable across those two comparisons.

5 Two concepts of rationality

5.1 Formal versus substantive rationality

In economics, the concept of rationality is usually interpreted in 'thin', formal terms. An agent is said to be rational, roughly, if his or her preferences and/or choices satisfy certain formal consistency constraints, and his or her beliefs are responsive to information in a Bayesian manner. While there are many ways of making the definition formally precise, practically all definitions of rationality in economics can be viewed as explications of this basic idea. In ordinary discourse as well as in philosophy, by contrast, we often employ the concept of rationality in a 'thicker', more substantive way, to imply something not only about the formal *consistency* of an agent's attitudes (preferences and beliefs), but also about their *content*. For instance, we often describe someone with self-destructive or otherwise 'unreasonable' preferences as 'irrational', even if those preferences and the resulting behaviour are internally consistent. The standard interpretation of rationality in economics would not licence this use of the term.

5.2 Hume versus Kant

Historically, the distinction between formal and substantive concepts of rationality is reflected nicely in the contrast between David Hume's and Immanuel Kant's ways of thinking about the requirements of rationality, which they called 'reason'. Like modern economic theory, Hume endorses a thin, formal conception of rationality, whereas Kant defends a much thicker, substantive conception. The following is an illustrative quote from Hume's *Treatise of Human Nature*:

> 'It is not contrary to reason [rationality in modern terms] to prefer the destruction of the whole world to the scratching of my finger. It is not contrary to reason for me to choose my total ruin... It is as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater... In short, a passion must be accompany'd with some false judgement [belief in modern terms], in order to its being unreasonable; and even then it is not the passion, properly speaking, which is unreasonable, but the judgment.' (Hume 1739, bk. 2, pt. 3, sect. 3)[5]

Kant, on the other hand, stresses that there are two kinds of rationality requirements, which he calls the 'hypothetical' and 'categorical imperatives' (Kant 1788). A 'hypothetical imperative' evaluates merely whether, *given the agent's ends*, the agent's means are effective in achieving those ends. This leaves open the question of whether the ends themselves are 'worthy' ones. In modern terms, the focus is solely on whether the agent's actions and choices are consistent with his or her preferences, not on whether those preferences are reasonable. This corresponds, once again, to the thin conception of rationality underlying modern economics. However, Kant proposes a 'categorical imperative' to evaluate the ends themselves. In modern terms, this requirement focuses not just on the internal consistency of the agent's preferences and choices, but also on their content, and here Kant's criterion is famously the universalizability of the agent's ends. But what matters for the purposes of this paper is not Kant's own criterion for evaluating the contents of an agent's preferences, but the distinction between formal and substantive criteria of rationality. The failure to be clear about this distinction tends to generate frequent misunderstandings between economists and philosophers (as well as people not trained in either discipline). We suggest that our property-based account of preference formation provides us with the formal resources to capture this distinction and to express different substantive, and not just formal, accounts of rationality.

5.3 Formalizing substantive accounts of rationality

We can obviously express the standard formal constraints of rationality in our model, and add to them the formal constraints given by our axioms. But we are also able to formalize two kinds of substantive constraints, each of which can in principle be of a prudential or of a moral kind:

(a) constraints on the normatively admissible weighing relations over property combinations, and
(b) constraints on the normatively relevant properties of the alternatives.

With regard to (a), we can ask whether the agent's actual weighing relation over property combinations meets the given normative constraints, i.e., whether the agent weighs different properties in a normatively admissible manner. And with regard to (b),

[5] For easier readability, Hume's expressions "'tis" and "chuse" have been replaced by the more modern forms "it is" and "choose".

we can compare

(i) the agent's motivation function from contexts to sets of motivationally salient properties, which captures his or her *actual* motivational dispositions

with (ii) an ideal function from contexts to sets of normatively admissible properties, which captures the properties the agent *ought* to focus on in forming his or her preferences in any context, according to some normative criterion.

This corresponds to the distinction between the agent's actual motivating reasons (or motives) for his or her preferences, and what the right normative reasons would be, given the relevant normative criterion. Using this distinction, we are then able to explore the relationship between actual and normatively ideal preferences.

In sum, introducing constraints under (a) and (b) allows us to distinguish, on the one hand, between preferences based on an admissible weighing relation and preferences based on an inadmissible one, and on the other hand, between preferences held for the right reasons and preferences held for the wrong reasons. Being able to draw these distinctions is an important feature of any substantive account of rationality as well as morality.

5.4 Some examples of substantive accounts

It is worth sketching some concrete examples of substantive accounts of rationality, including those that introduce moral motivations. A familiar substantive theory of rationality is the *self-interest theory*. According to it, the only normatively relevant properties of the alternatives are those that directly affect the agent in question. If the alternatives are allocations of goods, for example, then only properties of the i^{th} component of any allocation vector are deemed relevant to agent i. The weighing relation typically encodes some kind of 'more is better' principle. Another substantive theory, albeit a moral one, is a *utilitarian* theory. Here the normatively relevant properties of the alternatives are those that describe the happiness or welfare of any affected agent. In allocations, these are properties pertaining to all components, not just those corresponding to agent i. The weighing relation then takes some kind of additive form, whereby one property combination is ranked above another whenever the sum-total of the ascribed welfare in the first combination exceeds that in the second. A third illustrative theory, again of a moral kind, is a *Rawlsian* one. Under this theory, the normatively relevant properties of the alternatives are those that specify the level of primary goods and other resources held by the least advantaged members of the relevant society. The weighing relation then ranks one property combination above another whenever the ascribed level of goods or resources in the first combination exceeds that in the second. Interestingly, any positional dictatorship, including maximin and maximax, can be defined in terms of the same weighing relation, by specifying a different set of normatively relevant properties. It should be clear that many other normative theories, whether of a prudential or of a moral kind, can be expressed in our model.

While decision theory in the tradition of von Neumann and Morgenstern (1944), Savage (1954) and Jeffrey (1965/1983) offers a purely formal theory of rationality, the *homo economicus* thesis, which goes back at least to Adam Smith's *Wealth of Nations* (1776) and is still influential in many branches of economics, entails the conjunction of a formal theory and a substantive one. Its formal part coincides with standard decision

theory, but it adds to this a self-interest theory of human motivation (though Smith himself did not use the term *homo economicus* and acknowledges other motivations in *The Theory of Moral Sentiments* in 1759). So, ironically, even many economists, for instance in the areas of public choice and political economy, endorse a substantive theory of rationality, contrary to the official doctrine of defining rationality in thin, formal terms alone.

6 An alternative perspective on our theory

We have emphasized the idea that an agent's preference order over a given set of alternatives depends on the properties of the alternatives that are motivationally salient for the agent. The stable characteristic of the agent is no longer his or her preference order, but the underlying weighing relation over property combinations. No doubt, many economists will be reluctant to accept this departure from the standard theory, even if they agree that more needs to be said about how choices depends on contextual factors. Instead, they may try to explain this context-dependency without giving up the assumption of stable preferences. Unlike critics who view this assumption as restrictive and unrealistic, rational choice theorists see it as a virtue of their theory, which underlies its elegance and parsimony.

This raises the question of whether we could explain the phenomena captured by our theory in a more classical manner. One strategy would be:

- to introduce a sufficiently fine-grained 'ontology' of alternatives over which the agent would be assumed to hold fixed preferences[6] (this would be a refinement of the set X assumed in our model), and
- to reinterpret any preference change over the original, non-refined alternatives as an instance of an ordinary information-driven preference change over uncertain prospects, consistent with stable preferences over the refined alternatives.

On this strategy, the fundamental alternatives prior to the refinement would correspond to uncertain prospects in the refined ontology. What initially appeared to be a *fundamental* preference change would then be a change in the agent's *derived* preferences over non-fundamental prospects, driven by new information about their likely consequences. But although some preference changes might be explained in this manner, we do not think this strategy works in general. The reasons are formal and interpretational, and we here sketch them only briefly.

Formally, the dynamics of preference change consistent with this classical picture would be very different from the dynamics permitted by our model. First, any preference change would have to satisfy the constraints of Bayesian information learning. A preference reversal between two alternatives would be possible, roughly speaking, only if the agent came to assign lower subjective probabilities to the favourable consequences of the one alternative, relative to the favourable consequences of the other, and if that reassignment of probabilities respected Bayes's rule. Among other things, the agent would have had to assign non-zero probabilities to all relevant refined consequences; he or she could not have been unaware of some of them. Secondly, if all preference changes were information-driven, the agent would always be dynamically consistent. We would

[6] To be precise, we would need to specify those preferences by von Neumann-Morgenstern utilities.

not be able to account for any dynamic inconsistencies due to changes in preference. Thirdly, under the informational picture, many preference changes would be irreversible, as Bayesian information learning would always narrow down the set of alternatives to which the agent assigns non-zero probability. Unless we are willing to admit a combinatorial explosion of the set of fundamental alternatives, the possibility that someone might repeatedly switch back and forth between different preferences, depending on the context, would not be explicable.

Interpretationally, the main cost of remodelling every preference change in informational terms would be a significant expansion of the ontology of alternatives over which the agent would have to hold beliefs and preferences. This is a cognitively demanding model of an agent, which does not seem to be psychologically plausible. We would preserve rational choice theory's parsimony with respect to the assumption of fixed preferences only at the expense of sacrificing parsimony with respect to the cognitive complexity ascribed to the agent.

However, there is an alternative strategy by which we could accommodate the content of our theory while preserving the assumption of stable preferences. The strategy is to reinterpret the agent's weighing relation as a preference relation of a more fundamental kind, while introducing a distinction between the agent's ontology and that of the modeller. This distinction captures the idea that different contexts give rise to different 'lenses' through which the agent perceives the world. Suppose that, over and above the 'objective' set of alternatives X as described by the modeller, there exists a 'subjective' set of alternatives $\mathcal{X}$, which are the possible alternatives in the agent's perception. Specifically, each context $C \in \mathcal{C}$ gives rise to a *perception function*,

$$p_C : X \to \mathcal{X},$$

which assigns to each objective alternative $x \in X$ a corresonding subjective alternative $p_C(x) \in \mathcal{X}$, interpreted as the alternative x *as perceived by the agent in context* C.

We then assume that the agent's preferences over the subjective alternatives are fundamental and stable, while his or her preferences over the objective alternatives are derived and context-dependent. We can introduce a binary relation $\geq$ to represent the agent's stable preferences over $\mathcal{X}$, and a family of binary relations $(\succsim_C)_{C \in \mathcal{C}}$ to represent the agent's context-dependent preferences over X. Specifically, the agent prefers an objective alternative x over another, y, in context C if and only if he or she fundamentally prefers the subjective alternative $p_C(x)$ to the subjective alternative $p_C(y)$. Formally, for any $x, y \in X$ and any $C \in \mathcal{C}$,

$$x \succsim_C y \Leftrightarrow p_C(x) \geq p_C(y). \tag{1}$$

For example, if the agent perceives an objective alternative x solely as benefitting him- or herself while he or she perceives another, y, as less personally beneficial, then this will naturally lead to a preference for x over y. But if the agent perceives x as adversely affecting other people and y as having fewer negative externalities, then he or she may well arrive at the reverse preference, consistently with the same underlying preferences over subjective alternatives. Similarly, if the agent's perception function maps two distinct objective alternatives to the same subjective one, then it is natural for the agent to be indifferent between them. If someone perceives a café latte and an Australian flat white as the same thing, for instance, he or she will naturally be

indifferent between them. But if the agent's perception function changes and the two objective alternatives are mapped to distinct subjective ones, then the same underlying preference relation may well rank one of them above the other.

It should be evident that our property-based account of preference formation can be re-expressed in these terms. Here the subjective alternatives in $\mathcal{X}$ take the form of property combinations, and the agent's perception function maps each objective alternative to its set of motivationally salient properties in any given context. Formally, for $x \in X$ and any $C \in \mathcal{C}$,

$$p_C(x) = \{P \in M : x \text{ satisfies } P\}, \text{ where } M = f(C). \tag{2}$$

Substituting (2) into (1), we obtain the now-familiar structure of property-based preferences, i.e., for any $x, y \in X$ and any $C \in \mathcal{C}$,

$$x \succsim_C y \Leftrightarrow \{P \in M{:}x \,\text{satisfies}\, P\} \geq \{P \in M{:}y \,\text{satisfies}\, P\}, \text{where } M{=}f(C).$$

The weighing relation is then reinterpreted as a fundamental preference relation over subjective alternatives.

Whether the original interpretation of our model or this new, 'double-ontology' interpretation is more plausible is, to some extent, in the eye of the beholder. On the first interpretation, the agent's ontology of alternatives is objective and fixed, but preferences depend on the context and specifically on the agent's motivational state. The agent's stable characteristic is the underlying weighing relation, which, on this interpretation, is distinct from a preference relation. On the second interpretation, the agent's ontology of alternatives is subjective and variable, but we can reinterpret the agent's weighing relation as a stable preference relation over subjective alternatives. The difference between the two interpretations lies in the psychological account they give of the agent, and for this reason psychology may ultimately have to adjudicate between them.

7 An illustrative game-theoretic application

To illustrate how our model of preference formation can be used in standard game-theoretic applications, consider a simple two-player game whose form and material payoffs are those of the prisoners' dilemma, as shown in Table 1.[7]

	Cooperate	Defect
Cooperate	2,2	0,3
Defect	3,0	1,1

Table 1: A prisoners' dilemma in material payoffs

Consider the set X of possible outcomes (action pairs) of the game. Many properties of the elements of X might be of interest, for instance what the payoff of a player is,

[7] For earlier works on other-regarding feelings in games, such as sympathy and reciprocity but also spitefulness, see, e.g., Rabin (1993), Fehr and Gächter (1998), Bolton and Ockenfels (2000), Sethi and Somanathan (2001), Dufwenberg and Kirchsteiger (2004).

whether the resulting payoffs are Pareto-optimal, whether the distribution is equal or unequal, and so on. For simplicity, let us focus on properties of the following kind:

$$\text{`}i \text{ gets } m\text{'},$$

where i is a player and m a possible material payoff (e.g., amount of money). A self-interested player will be motivated only by properties of the alternatives that affect him- or herself. Thus, if player i is self-interested, his or her motivational state will be given by the set of properties

$$M = \{\text{`}i \text{ gets } m\text{'} : m \text{ is a possible payoff}\}.$$

By contrast, an other-regarding player will also be motivated by properties of the alternatives that affect other players. If player i is other-regarding, his or her motivational state will be given by the set of properties

$$M = \{\text{`}j \text{ gets } m\text{'} : m \text{ is a possible payoff, } j \text{ is a player}\}.$$

Clearly, self-interested and other-regarding players perceive the alternatives in Table 1 differently. For a self-interested player 1, the sets of motivationally salient properties of the four possible outcomes are simply {'1 gets 2'}, {'1 gets 3'}, {'1 gets 0'}, and {'1 gets 1'}, while for an other-regarding player 1, they are {'1 gets 2', '2 gets 2'}, {'1 gets 3', '2 gets 0'}, {'1 gets 0', '2 gets 3'}, and {'1 gets 1', '2 gets 1'}. The case for player 2 is analogous.

To keep the example simple, suppose that any player's weighing relation $\geq$ ranks different property combinations in terms of the sum-total of the payoffs listed in them, formally, for any S_1, S_2,

$$S_1 \geq S_2 \Leftrightarrow W(S_1) \geq W(S_2),$$

where the weight of any property combination S is

$$W(S) = \sum_{m,j \,:\, \text{`}j \text{ gets } m\text{'} \in S} m.$$

For example, the weight of the property combination {'1 gets 2'} is 2, while that of the property combination {'1 gets 2', '2 gets 2'} is 4, and so on.

Given this weighing relation, the players' preferences over the four possible outcomes are now straightforwardly induced by their motivational state (self-interested or other-regarding), as shown in Figure 1. Thus the material game form introduced above induces four different games in preference terms, with four different Nash equilibria, as underlined in each matrix. The resulting behavioural prediction depends crucially on the players' motivational states, and thereby on our hypothesis about which contexts trigger which states. Notably, for other-regarding players, the prisoners' dilemma *in material terms* is not a prisoners' dilemma *in preference terms*: to the contrary, cooperation becomes a dominant strategy for any such player.

As this simple example shows, our model can systematically describe the mechanism by which a material game form is transformed into a fully specified game, depending on which properties of the outcomes are rendered motivationally salient for the players. This, in turn, provides us with a basis for studying endogenous preferences in games more generally.

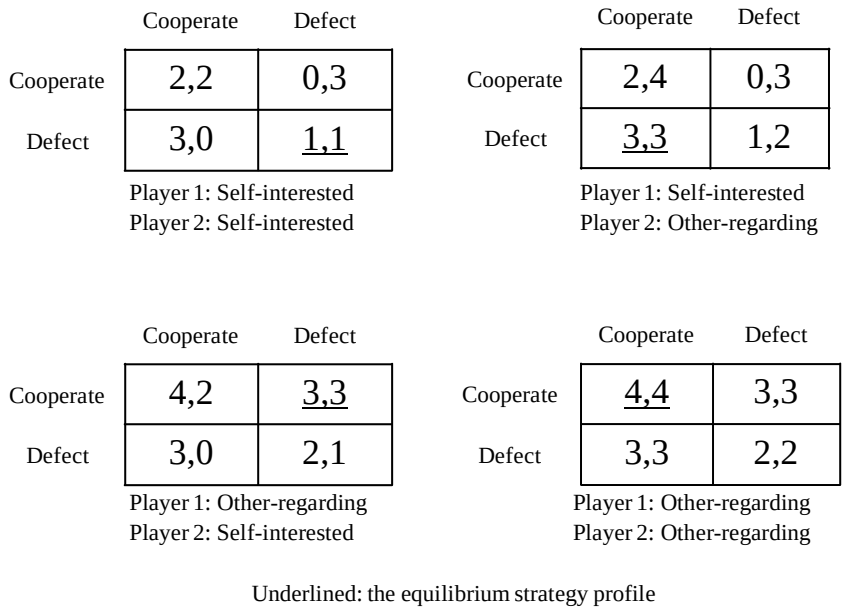

Figure 1: The players' preferences in different motivational states

8 Concluding remarks

Our aim has been to connect two distinct ways of thinking about an agent's preferences:

- Economists tend to follow the classical instrumental model of rational agency that goes back to David Hume. The model's strength is its parsimony, but its weakness is its inability to account for preference formation or genuine preference change.
- Philosophers and others tend to be interested in a more substantive model of rational agency, under which we can account for the motivations behind an agent's preferences and for genuine preference change, and under which we can normatively assess the content of those preferences.

By supplementing standard rational choice theory with a property-based account of preference formation, our proposal seeks to build a bridge between these two ways of thinking.

9 References

Bolton, G. E., Ockenfels, A. (2000) ERC: a theory of equity, reciprocity and competition. *American Economic Review* 90: 166-193.

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

Bradley, R. (2007) The Kinematics of Belief and Desire. *Synthese* 56(3): 513-535.

Dekel, E., Ely, J., Yilankaya, O. (2007) Evolution of preferences. *Review of Economic Studies* 74: 685-704.

Dietrich, F. (2008a) Anti-terrorism politics and the risk of provoking. Working

paper, London School of Economics.

Dietrich, F. (2008b) Modelling change in individual characteristics: an axiomatic approach. Working paper, London School of Economics.

Dietrich, F., List, C. (2010) A reason-based theory of rational choice. Working paper, London School of Economics.

Dietrich, F., List, C. (forthcoming) A model of non-informational preference change. *Journal of Theoretical Politics*.

Dryzek, J., List, C. (2003) Social Choice Theory and Deliberative Democracy: A Reconciliation. *British Journal of Political Science* 33(1): 1-28.

Dufwenberg, M., Kirchsteiger, G. (2004) A theory of sequential reciprocity. *Games and Economic Behavior* 47(2): 268-298.

Fehr, E., Gächter, S. (1998) Reciprocity and economics: the economic implications of homo reciprocans. *European Economic Review* 42: 845-859.

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

Hammond, P. (1976) Changing tastes and coherent dynamic choice. *Review of Economic Studies* 43: 159-173.

Hume, D. (1739) *A Treatise of Human Nature*. Reprinted from the original edition and edited by Selby-Bigge, L. A. (1896). Oxford (Clarendon Press).

Jeffrey, R. (1965/1983) *The Logic of Decision*. Chicago (University of Chicago Press).

Kant, I. (1788) *Critique of Practical Reason*. Translated by Beck, L. W. (1949). Chicago (University of Chicago Press).

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

O'Donoghue, E., Rabin, M. (1999) Doing it now or doing it later. *American Economic Review* 89: 103-124.

Pettit, P. (1991) Decision theory and folk psychology. In Bacharach, M., Hurley, S. (eds.) *Foundations of Decision Theory*. Oxford (Blackwell): 147-175.

Polak, R. (1976) Interdependent preferences. *American Economic Review* 66(3): 309-320.

Rabin, M. (1993) Incorporating fairness into game theory and economics, Amer. Econ. Rev. 83: 1281-1302.

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

Titmuss, R. M. (1970) *The gift relationship: From human blood to social policy*. London (Allen and Unwin).

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

Sethi, R., Somanathan, E. (2001) Preference evolution and reciprocity. *Journal of Economic Theory* 97: 273-297.

Smith, A. (1759) *The Theory of Moral Sentiments*. Edited by Haakonssen, K. (2002). Cambridge (Cambridge University Press).

Smith, A. (1776) *An Inquiry into the Nature and Causes of the Wealth of Nations.* Edited by Sutherland, K. (2008). Oxford (Oxford University Press).

Strotz, R. H. (1955-56) Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies* 23(3): 165-180.

Tversky, A., Kahneman, D. (1981) The framing of decisions and the psychology of choice. *Science* 211(4481): 453-458.

von Neumann, J., Morgenstern, O. (1944) *Theory of Games and Economic Behaviour.* Princeton (Princeton University Press).

A Appendix: proofs

Notation. Let us write $\mathcal{P}$ to denote the set of all relevant properties (e.g., we could stipulate $\mathcal{P} := \bigcup_{M \in \mathcal{M}} M$). Each property (e.g., $P, Q, ... \in \mathcal{P}$) defines a set of alternatives with that property, called the *extension* of the property and denoted by putting the property symbol in bold (e.g., $\mathbf{P}, \mathbf{Q}... \subseteq X$). The set of properties of x ($\in X$) that belong to M ($\subseteq \mathcal{P}$) is denoted by $M_x = \{P \in M : x \text{ satisfies } P\} = \{P \in M : x \in \mathbf{P}\}$. The set of all property combinations (consistent sets of properties) is denoted by $\mathcal{S} = \{S \subseteq \mathcal{P} : \cap_{P \in S} \mathbf{P} \neq \varnothing\}$.

We begin with a lemma, which is useful to prove both theorems.

Lemma 1 *Suppose Axiom 1 holds. For all $x, y, x', y' \in X$ and all $M \in \mathcal{M}$, if $M_x = M_{x'}$ and $M_y = M_{y'}$ then $x \succsim_M y \Leftrightarrow x' \succsim_M y'$.*

Proof. For $x, y, x', y' \in X$ and $M \in \mathcal{M}$, if $M_x = M_{x'}$ and $M_y = M_{y'}$, then, under Axiom 1, $x \sim_M x'$ and $y \sim_M y'$, whence by transitivity $x \succsim_M y \Leftrightarrow x' \succsim_M y'$. ■

The following proof of Theorem 1 is closely analogous to that of the first theorem in Dietrich and List (2009), but applies more generally since the present paper explicitly treats properties intensionally rather than extensionally, distinguishing between a property $P \in \mathcal{P}$ and its extension $\mathbf{P} \subseteq X$.

Proof of Theorem 1. First consider the easy direction of implication. If a binary relation $\geq$ on $\mathcal{S}$ represents the family of preference orders $(\succsim_M)_{M \in \mathcal{M}}$, then Axiom 2 holds obviously. To see that Axiom 1 holds, consider $M \in \mathcal{M}$ and $x, y \in X$ such that $M_x = M_y$. As $\succsim_M$ is reflexive, $x \sim_M x$, whence $M_x \equiv M_x$ since $\geq$ induces $\succsim_M$. Since $M_x \equiv M_x$ and $M_x = M_y$, it follows that $M_x \equiv M_y$. So, $x \sim_M y$, again since $\geq$ induces $\succsim_M$.

Now consider the non-trivial direction. Suppose Axioms 1 and 2 hold, and suppose $\mathcal{M}$ is intersection-closed (an assumption not needed above for the easy direction).

Claim 1. For all $x, y, x', y' \in X$ and all $M, M' \in \mathcal{M}$, if $M_x = M'_{x'}$ and $M_y = M'_{y'}$, then $x \succsim_M y \Leftrightarrow x' \succsim_{M'} y'$.

Consider any $x, y, x', y' \in X$ and $M, M' \in \mathcal{M}$ such that $M_x = M'_{x'}$ and $M_y = M'_{y'}$. As $\mathcal{M}$ is intersection-closed, $M \cap M' \in \mathcal{M}$. Now

$$(M \cap M')_x = (M \cap M')_{x'} = M_x = M'_{x'},$$

because, firstly, $M_x = M'_{x'}$; secondly, $(M \cap M')_x = M_x$, as $(M \cap M')_x = M_x \cap M'_x = M_x$ (the last identity holds because $M'_x \supseteq (M'_{x'})_x = (M_x)_x = M_x$); and, thirdly, $(M \cap M')_{x'} = M'_{x'}$, as $(M \cap M')_{x'} = M_{x'} \cap M'_{x'} = M'_{x'}$ (the last identity holds because $M_{x'} \supseteq (M_x)_{x'} = (M'_{x'})_{x'} = M'_{x'}$). Analogously,

$$(M \cap M')_y = (M \cap M')_{y'} = M_y = M'_{y'}.$$

Since $(M \cap M')_x = M_x$ and $(M \cap M')_y = M_y$, Axiom 2 implies

$$x \succsim_{M \cap M'} y \Leftrightarrow x \succsim_M y. \tag{3}$$

Further, since $(M \cap M')_{x'} = M'_{x'}$ and $(M \cap M')_{y'} = M'_{y'}$, Axiom 2 implies

$$x' \succsim_{M \cap M'} y' \Leftrightarrow x' \succsim_{M'} y'. \tag{4}$$

Finally, since $(M \cap M')_x = (M \cap M')_{x'}$ and $(M \cap M')_y = (M \cap M')_{y'}$, Lemma 1 implies

$$x \succsim_{M \cap M'} y \Leftrightarrow x' \succsim_{M \cap M'} y'. \tag{5}$$

Combining the equivalences (3), (4) and (5), $x \succsim_M y \Leftrightarrow x' \succsim_{M'} y'$. ■

Claim 1 allows us naturally to define a binary relation $\geq$ on $\mathcal{S}$: for all $S, S' \in \mathcal{S}$, $S \geq S'$ if and only if $x \succsim_M y$ for some (hence, by Claim 1, *all*) $x, y \in X$ and $M \in \mathcal{M}$ such that $M_x = S$ and $M_y = S'$.

Claim 2. For every $M \in \mathcal{M}$, $\geq$ induces $\succsim_M$, i.e., $x \succsim_M y \Leftrightarrow M_x \geq M_y$ for all $x, y \in X$.

Consider any $M \in \mathcal{M}$ and $x, y \in X$. Suppose first that $x \succsim_M y$. To show that $M_x \geq M_y$, we need to find $x', y' \in X$ and $M' \in \mathcal{M}$ such that $M'_{x'} = M_x$, $M'_{y'} = M_y$ and $x' \succsim_{M'} y'$. Simply take $x' = x$, $y' = y$, and $M' = M$. Conversely, assume that $M_x \geq M_y$. By the definition of $\geq$ and Claim 1, we have $x' \succsim_{M'} y'$ for *all* $x', y' \in X$ and $M' \in \mathcal{M}$ such that $M'_{x'} = M_x$ and $M'_{y'} = M_y$. So, in particular $x \succsim_M y$, which completes the proof. ■

We now turn to the proof of Theorem 2, drawing on Theorem 1.

Proof of Theorem 2. First, if a separable weighing relation $\geq$ represents the family of preference orders $(\succsim_M)_{M \in \mathcal{M}}$, then Axiom 3 holds obviously, and Axiom 1 holds for the same reason as the one given in the proof of Theorem 1.

Now assume that Axioms 1 and 3 hold, and $\mathcal{M}$ is subset-closed. Since Axiom 3 implies Axiom 2 and subset-closure implies intersection-closure, we know from Theorem 1 that there is a weighing relation $\geq$ which represents the family $(\succsim_M)_{M \in \mathcal{M}}$. This relation is not generally separable[8] and can therefore not ultimately be used to establish that preferences are property-based in a separable way. Nonetheless, we start by considering $\geq$. Call property combinations $S, S' \in \mathcal{S}$ *ranked* by $\geq$ if $S \geq S'$ or $S' \geq S$. We can assume without loss of generality that $\geq$ ranks only pairs of sets $S, S' \in \mathcal{S}$ which feature

[8] E.g., Axioms 1 and 2 do not generally rule out the existence of a property combination $S \in \mathcal{S}$ such that $S \not\geq S$. If $\geq$ were separable, it would follow that $\varnothing \not\geq \varnothing$. But for all $x \in X$, we have $x \succsim_M x$ where $M = \varnothing$, and hence $M_x \geq M_x$, i.e., $\varnothing \geq \varnothing$ (note that $\varnothing \in \mathcal{M}$ as $\mathcal{M}$ is subset-closed).

in the representation of preferences, i.e., which are of the form $S = M_x$ and $S' = M_{x'}$ for some $M \in \mathcal{M}$ and $x, x' \in X$.

Claim 1. If two sets $S, S' \in \mathcal{S}$ are ranked by $\geq$, then $S \geq S' \Leftrightarrow S\backslash C \geq S'\backslash C$ for all $C \subseteq S \cap S'$. (So, $\geq$ is separable in a restricted sense.)

Consider any sets $S, S' \in \mathcal{S}$ ranked by $\geq$ and any subset $C \subseteq S \cap S'$. As S and S' are ranked by $\geq$, there are $M \in \mathcal{M}$ and $x, x' \in X$ such that $S = M_x$ and $S' = M_{x'}$. As $\mathcal{M}$ contains M and is subset-closed, $\mathcal{M}$ also contains the set $M^* := M\backslash C$. Since $M\backslash M^* = C \subseteq S \cap S' = M_x \cap M_{x'}$, all properties in $M\backslash M^*$ are satisfied by both x and x'. So, by Axiom 3, $x \succsim_M x' \Leftrightarrow x \succsim_{M^*} x'$. Hence, $M_x \geq M_{x'} \Leftrightarrow M^*_x \geq M^*_{x'}$, as $\geq$ represents the preferences. In other words, $S \geq S' \Leftrightarrow S\backslash C \geq S'\backslash C$. ■

We now define another weighing relation from $\geq$, to be denoted $\geq^*$, which represents the agent's preferences (Claim 2) and is separable (Claim 3). For all $S, S' \in \mathcal{S}$, we define $S \geq^* S'$ if and only if $S\backslash C \geq S'\backslash C$ for $C = S \cap S'$; hence, if and only if $S\backslash S' \geq S'\backslash S$.

Claim 2. $\geq^*$ agrees with $\geq$ on all pairs ranked by $\geq$ (i.e., $S \geq^* S' \Leftrightarrow S \geq S'$ for all $S, S' \in \mathcal{S}$ ranked by $\geq$), whence $\geq^*$ still induces each preference order $\succsim_M$, $M \in \mathcal{M}$.

Suppose $S, S' \in \mathcal{S}$ are ranked by $\geq$. By definition, $S \geq^* S'$ if and only if $S\backslash C \geq S'\backslash C$ where $C = S \cap S'$. By Claim 1, the latter holds if and only if $S \geq S'$. ■

Claim 3. $\geq^*$ is separable (which completes the proof).

We consider any $S, S' \in \mathcal{S}$ and any $C \subseteq S \cap S'$, and show that $S \geq^* S' \Leftrightarrow S\backslash C \geq^* S'\backslash C$ (implying that $\geq^*$ is separable). By definition of $\geq^*$, the expression on the left-hand side of this equivalence means that $S\backslash S' \geq S'\backslash S$, while the expression on the right-hand side means that $(S\backslash C)\backslash(S'\backslash C) \geq (S'\backslash C)\backslash(S\backslash C)$. So, noting that $(S\backslash C)\backslash(S'\backslash C) = S\backslash S'$ and $(S'\backslash C)\backslash(S\backslash C) = S'\backslash S$, the two expressions mean the same, hence are equivalent. ■

We finally give an example showing that Theorem 2 would not hold if $\mathcal{M}$ were merely assumed to be intersection-closed (as in Theorem 1) and not subset-closed. Suppose $\mathcal{M}$ contains just one set M ($\neq \varnothing$), so that $\mathcal{M}$ is trivially intersection-closed but not subset-closed. Consider a weighing relation $\geq$ which is a non-separable weak order over property combinations. Assume $\succsim_M$ is the preference order generated by $\geq$, so that $x \succsim_M y \Leftrightarrow M_x \geq M_y$ for all $x, y \in X$. Axioms 1 holds as $\geq$ is reflexive, and Axiom 3 holds trivially. Yet the weighing relation $\geq$ is non-separable, and under mild additional conditions any other weighing relation is also non-separable. As an extreme case in which every weighing relation is non-separable, suppose $M = \mathcal{P}$, and the properties are mutually independent, i.e., for every property set $S \subseteq \mathcal{P}$ there is an alternative of which each property in S but none in $\mathcal{P}\backslash S$ is true. Then the weighing relation is uniquely determined by $\succsim_M$, so that every weighing relation – there is only one, namely $\geq$ – is non-separable.