A reason-based theory of rational choice[*]

Franz Dietrich and Christian List

London School of Economics

29 October 2009, revised 23 April 2010

Abstract

There is a surprising disconnect between formal rational choice theory and philosophical work on reasons. The one is silent on the role of reasons in rational choices, the other rarely engages with the formal models of decision problems used by social scientists. To bridge this gap, we propose a new, reason-based theory of rational choice. At its core is an account of preference formation, according to which an agent's preferences are determined by his or her motivating reasons, together with a 'weighing relation' between different combinations of reasons. By explaining how someone's preferences may vary with changes in his or her motivating reasons, our theory illuminates the relationship between deliberation about reasons and rational choices. Although primarily positive, the theory can also help us think about how those preferences and choices ought to respond to normative reasons.

1 Introduction

The idea that a rational choice is a choice based on reasons and that a rational agent is someone who acts on the basis of reasons – at least those reasons the agent takes him- or herself to have – is a very natural one, and yet reasons are largely absent from modern rational choice theory. Instead, rational choice theory, also known as decision theory, is paradigmatically Humean. A rational agent, on the standard picture, has beliefs and desires (typically modelled as assignments of probabilities and utilities to different possible worlds, states or outcomes), and acts so as to satisfy his or her desires in accordance with his or her beliefs.[1] On this picture, the agent's desires over

---

[1]Classic contributions include John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1944), Leonard Savage, *The Foundations*

possible worlds or fully specified outcomes – his or her *fundamental preferences* – are volitional attitudes that are entirely separate from, and do not respond to, the agent's beliefs, which are cognitive attitudes. Only desires over prospects with uncertain outcomes can change in response to belief changes about the relative likelihood of these outcomes; these 'surface-level' desires are also called *derived preferences*. The assumption of fixed fundamental preferences outside the realm of rational scrutiny contrasts with the more common view of agency as involving the capacity to form, revise and rationally pursue one's conception of the good, as Rawls famously describes it.[2] Standard rational choice theory focuses exclusively on the rational pursuit of an agent's preferences, and is silent on how these preferences are formed and how they may be revised, for instance by deliberating about and responding to various reasons.

The aim of this paper is to present an alternative theory of rational choice, which gives reasons their proper place. Of course, there is a large body of philosophical work on the relationship between reasons and actions.[3] But there is currently no *formal* theory of rational choice that is reason-based.[4] As a result, decision theorists and social scientists engaged in the formal modelling of decision problems lack the conceptual resources for capturing the role played by reasons in rational decision making. Similarly, some important philosophical debates about reasons have not yet been cast in

of Statistics (New York: Wiley, 1954), and Richard Jeffrey, *The Logic of Decision* (Chicago: University of Chicago Press, 1965/1983). Savage defines probabilities over states of the world, utilities over outcomes of actions in them; Jeffrey defines both probabilities and utilities over possible worlds.

[2] See, e.g., John Rawls, *Political Liberalism* (New York: Columbia University Press, 1993).

[3] Important works include Derek Parfit, *Reasons and Persons* (New York: Oxford University Press, 1984), Thomas Scanlon, *What We Owe to Each Other* (Cambridge, MA: Harvard University Press, 1998), Joseph Raz, *Practical Reason and Norms* (Oxford: Oxford University Press, 1999), and the essays in R. Jay Wallace, Philip Pettit, Samuel Scheffler and Michael Smith (eds.), *Reason and value: themes from the moral philosophy of Joseph Raz* (Oxford: Oxford University Press, 2004), notably John Broome's chapter, 'Reasons', pp. 28-55. See also the surveys by Stephen Finlay and Mark Schroeder, 'Reasons for Action: Internal vs. External' (2008), James Lenman, 'Reasons for Action: Justification vs. Explanation' (2009), and Michael Ridge, 'Reasons for Action: Agent-Neutral vs. Agent-Relative' (2005), in the *Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu/).

[4] Works pertaining to, but not framed in terms of, reasons include Ralph Keeney and Howard Raiffa's multi-attribute utility theory in *Decisions with Multiple Objectives* (Cambridge: Cambridge University Press, 1993) and Isaac Levi's work on unresolved value conflicts in *Hard Choices* (Cambridge: Cambridge University Press, 1986). In 'Rationalization' (University of Pennsylvania, 2008), Vadim Cherepanov, Timothy Feddersen and Alvaro Sandroni generalize standard rational choice theory to account for what they call 'rationalization': agents have fixed underlying preferences but make choices they can 'rationalize', e.g., justify in public. Their model remains close to standard rational choice theory in that the assumption of fixed fundamental preferences is not lifted. Relatedly, Paola Manzini and Marco Mariotti, in 'Sequentially Rationalizable Choice', *American Economic Review* 97 (5) (2007), pp. 1824-1839, model how human choices may result from the application of multiple choice heuristics. Some of our concerns are also shared by Ariel Rubinstein in his discussion in *Economics and Language* (Cambridge: Cambridge University Press, 2000) of how an agent's language may constrain his or her preferences, but he does not develop an account of reasons.

formal terms.[5] This is a significant gap in the literature, which we here seek to fill.

At the core of our theory is a reason-based account of preference formation, which can be roughly summarized as follows. An agent's preferences over the relevant fundamental objects (possible worlds, states, outcomes) depend on the reasons that motivate him or her and may vary with changes in them. A motivating reason, as we understand it, is a proposition that is motivationally relevant for the agent's preferences towards the objects of which it is true: it may affect those preferences. (We later contrast this with a normative reason, which is a proposition that is normatively relevant for those preferences, constraining the preferences the agent *ought* to have.) The relationship between motivating reasons and preferences is governed by two axioms, introduced below, which are necessary and sufficient for a parsimonious representation of those preferences across variations in the agent's motivational state.

The preferences across such variations are then representable in terms of a single *weighing relation*, a binary relation that ranks different possible combinations of reasons relative to one another. It may rank, for instance, the combination 'I have enough food' and 'I am healthy' above the combination 'I do not have enough food' and 'I am unhealthy'; or the combination 'the economy is growing' and 'there is no war' above the one in which only one of these reasons is present. The agent now prefers an alternative, say a particular state of the world, to another if and only if his or her weighing relation ranks the combination of motivating reasons true of the first alternative above the one true of the second. The two combinations of reasons that are being compared characterize the two alternatives from the perspective of the agent's motivational state. In this way, the agent's weighing relation, together with his or her motivating reasons, determines his or her preferences. The weighing relation can be interpreted in a number of ways, depending on one's philosophical vantage point; our analysis is deliberately ecumenical.[6]

Our theory is both ambitious and flexible. It not only captures the idea that an important property of a rational choice is its responsiveness to certain reasons, but it also shows us how changes in the set of reasons motivating an agent can lead to changes in his or her preferences, even at the level of possible worlds or fully specified outcomes, where standard rational choice theory denies such changes. Despite our focus on the relationship between motivating reasons and actual preferences, our

---

[5] In *Reason and Rationality* (Princeton: Princeton University Press, 2009), Jon Elster confirms this observation: 'Whereas the theory of rational choice has been elaborated and developed with great precision, the same cannot be said of the idea of reason' (p. 7).

[6] On a 'cognitivist' interpretation, the weighing relation encodes a set of judgments about the relative 'goodness' of different possible reason combinations; on a 'non-cognitivist' one, it encodes the agent's dispositions to prefer some such combinations to others. In the first case, the agent may be taken to judge, e.g., that the combination 'I have enough food' and 'I am healthy' is better than its propositionwise negation; in the second, he or she may simply be disposed to prefer an alternative instantiating the first combination to one instantiating the second if the reasons are each motivating.

theory can be reinterpreted to formalize the relationship between normative reasons and the preferences an agent ought to have, thereby shedding light on some normative questions.[7] By suitably varying the interpretation of 'reasons' and 'preferences', our theory can thus be put to positive as well as normative uses. In consequence, the theory offers novel resources for illuminating the relationship between deliberation about reasons and rational choices, which in turn is relevant to many philosophical and social-scientific questions. Furthermore, our theory generalizes standard rational choice theory, entailing it as a special case, and therefore pinpoints precisely in what sense the standard theory is at best incomplete.[8]

In Sections 2 and 3, we introduce some basic concepts – alternatives, preferences, and reasons – and discuss what it means for a reason to be motivating. In Sections 4 and 5, we explain how our theory depicts an agent's possible psychological states and introduce our two central axioms on the relationship between reasons and preferences. In Sections 6, 7 and 8, we present our main representation theorems and discuss their interpretation. With this core of the theory in place, we comment on some philosophical questions, first, in Section 9, on the distinction between reason-based explanation and reason-based justification and then, in Section 10, on the role of reasons in an agent's rational deliberation and on their relevance to the resolution of disagreements between different agents' preferences. Finally, in Sections 11 and 12, we add two technical extensions without which our theory would not be complete: we show how it can handle preferences under uncertainty and explain how an agent's reason-based preferences relate to his or her choices. Proofs are given in an appendix.

2   Alternatives, preferences, and reasons

We consider an agent's preferences over some fundamental objects of preference, which we call *alternatives*. Depending on the area of application, the alternatives could be, for example, possible worlds or states of the world, outcomes of actions, bundles of goods, policy programmes, or election options. What matters is that the alternatives are mutually exclusive and jointly exhaustive of the relevant space of possibilities. Later we also consider preferences over general prospects, that is, probability distributions over alternatives, so as to capture the fact that agents often do not choose between individual alternatives, but only between different uncertain prospects, which result from the actions they can take.

Let X denote the set of alternatives. The agent's preferences over the elements

---

[7] The 'ought' involved allows various interpretations, such as 'ideally rational' versus 'moral' ones, depending on what kinds of normative reasons we wish to capture.

[8] For a discussion of some limitations of standard rational choice theory, supporting our current perspective, see Philip Pettit, 'Decision Theory and Folk Psychology', in Michael Bacharach and Susan Hurley (eds.), *Foundations of Decision Theory* (Oxford: Blackwell, 1991), pp. 147-175. See further Pettit's collection, *Rules, Reasons, and Norms* (Oxford: Oxford University Press, 2002).

of X are represented by some order $\succsim$ on X.[9] For any two alternatives x and y, we write $x \succsim y$ to mean that the agent weakly prefers x to y. We further write $x \succ y$ if $x \succsim y$ but not $y \succsim x$ (a strict preference for x over y), and $x \sim y$ if $x \succsim y$ and $y \succsim x$ (an indifference between x and y). Later we make explicit the way these preferences affect the agent's choices or actions.

We are interested in how an agent's preferences depend on his or her reasons. There are a number of ways the concept of a reason might be formalized. Some philosophers think of reasons as certain kinds of facts; others as certain kinds of properties of the alternatives under consideration; still others as mental states of the agent. For our purposes, we model reasons as special kinds of propositions, as explained in a moment. To keep things simple, we define a *proposition* as a subset of X; it is said to be *true* of those alternatives contained in it, and *false* of all others.[10] But it is sometimes useful to represent propositions by sentences from a suitable language – especially when we want to express two or more distinct things that are true of the same set of alternatives in X – and our formal results continue to hold in this case as well.

Now a *reason* is a proposition that has a particular kind of relevance for the agent's preferences towards the alternatives of which it is true.[11] Depending on how we spell out that relevance, we obtain different conceptions of a reason. A *motivating reason* is a proposition that is motivationally relevant for the agent's preferences: if true of an alternative, it may affect the agent's *actual* preference for the given alternative vis-à-vis others. A *normative reason*, by contrast, is a proposition that is normatively relevant for those preferences: if true of an alternative, it may affect the preference the agent *ought* to have for that alternative vis-à-vis others, whether or not he or she actually has that preference. (The 'ought' can be interpreted in various ways, depending on the kind of normativity we wish to capture.) We deliberately leave open *how exactly* a proposition must affect or constrain the agent's actual or ideal preferences to count as a motivating or normative reason.[12] This may depend on a

---

[9] Formally, $\succsim$ is a complete and transitive binary relation on X.

[10] This definition is standard when X is a set of possible worlds or states of the world. When X consists of other objects (e.g., bundles of goods), it is more conventional to call a subset of X a *property*. But any property can be associated with a proposition: the proposition that an object has that property. This proposition is true of the objects that have the property, and false of all others. Bearing this in mind, we refer to propositions in general. This terminology has some independent advantages, including (but not restricted to) the representability of propositions by sentences.

[11] We thus focus on reasons for preferences, which then constrain choices or actions. Our analysis can be adjusted so as to formalize reasons for choices or actions without preferences as intermediaries, by translating our axioms below into constraints on the relationship between reasons and choice functions. Needless to say, the use of any philosophically loaded concept, such as that of a reason, in a formal theory requires some regimentation, which may not capture every established usage.

[12] In our definitions, it suffices to interpret 'may affect' as synonymous with 'is eligible to affect'. More restrictive definitions are obtained by strengthening the wording to 'makes a difference to'.

number of factors, including the context and which other reasons are present. Our characterization theorems below provide a precise treatment of these issues.

If someone decides to go to a café because of a craving for coffee, one of her motivating reasons is that coffee is available there. This proposition is motivationally relevant for the agent's preferences in that it affects her actual preference for the alternative of going to the café, of which the proposition is true. If someone prefers to drive recklessly despite endangering himself and others, the proposition that his driving is dangerous is a normative reason for him to give up this preference, even if he is not motivated by this consideration. The proposition is normatively relevant for his preferences in that it implies something for the preferences he *ought* to have: he ought to disprefer an alternative – reckless driving – of which the proposition is true. In what follows, we focus primarily on how an agent's actual preferences depend on his or her motivating reasons, although, as discussed later, our formal analysis can be reinterpreted so as to capture the way an agent's normative reasons constrain the preferences he or she *ought* to have.

We write M to denote the set of motivating reasons for the agent's preferences in a given psychological state. It contains all propositions that play a role in the agent's preference formation over the alternatives in X. Thus M need not be a consistent set: an agent can be simultaneously motivated by some reasons jointly true only of x and others jointly true only of y, where these pull in opposite directions and need to be weighed relative to one another. The set M simply captures the motivational state in which the agent forms his or her preferences over the alternatives. It is, in turn, a subset of some underlying set of possible reasons, which we call $\mathcal{P}$. Nothing much hinges on how permissively or restrictively we define $\mathcal{P}$, so long as it includes at least those propositions that *could* become motivationally relevant for the agent.[13] Examples of such propositions are 'there is war', 'there is food available', 'the dish is poisonous', 'I am hungry', 'the power station has high CO_2 emissions', and so on.

To indicate the dependency of the agent's preferences on his or her set of motivating reasons, we append the subscript M to the symbol $\succsim$, interpreting $\succsim_M$ as the agent's preference order in the event that M is his or her set of motivating reasons in relation to the alternatives in X. Further, $\succ_M$ represents the corresponding strict preference, and $\sim_M$ the indifference relation.

---

[13] The set $\mathcal{P}$ could be the set of all propositions, but it could also be much smaller. This allows us to capture a great variety of assumptions about which propositions are possible reasons. Our theory is consistent, e.g., with defining $\mathcal{P}$ according to some psychological account of which propositions can motivate someone, but also with including in $\mathcal{P}$ additional propositions that may serve as normative reasons but are unlikely to motivate a given agent. Our theory is further consistent with the view that $\mathcal{P}$ consists of all propositions that can be expressed by sentences of a particular form (e.g., sentences with, or without, certain predicates, operators, or connectives), and even with the view that $\mathcal{P}$ has certain closure properties (e.g., under conjunction or disjunction).

3 Which reasons are motivating?

From a psychological perspective, not every possible reason pertaining to the alternatives in X will become motivating for an agent in a given context. This depends very much on the agent's psychology and the context in question. And from a normative perspective, not every motivating reason might be deemed appropriate. A proposition's capacity to motivate someone is quite distinct from its normative relevance, and thus motivating reasons and normative reasons can come significantly apart. Someone might be psychologically motivated by reasons which, normatively speaking, we find utterly deplorable. Conversely, a given proposition might seem to be a compelling normative reason for or against something from a third-person perspective – or from the perspective of some background normative theory – and yet it may fail to motivate an agent. Since our primary aim is to develop a positive theory of rational choice, we focus on how an agent's motivating reasons explain his or her preferences and choices, independently of whether these are also normative reasons justifying them. Later, however, we return to the latter, normative concern.

There are many possible accounts of when a proposition attains motivational relevance for an agent's preferences. We need not commit ourselves to one such account here. According to a first, rather simple account, a proposition becomes motivating as soon as the agent conceptualizes it abstractly – in the sense that, in his or her conceptualization of the world in the relevant context, the agent distinguishes between those alternatives of which the proposition is true and those of which it isn't. If our conceptualization of the world does not distinguish between those states of the world in which the number of grains of sand is even and those in which it is odd, for example, then the proposition that there exists an even number of grains of sand cannot be a motivating reason that affects our rational choices.[14] However, while the agent's ability to conceptualize a given proposition seems necessary for it to gain motivational relevance, it may not be enough.

On a second account, a proposition becomes motivating only when the agent qualitatively – and not merely abstractly – understands it. A policy maker, for example, may abstractly understand that different foreign policies can be distinguished from each other with respect to whether or not they make cheap oil available and whether or not they lead to war, but fail to understand qualitatively what a war involves and thus fail to be motivated by the latter consideration. This account of how a proposition becomes motivating requires that the distinction between abstract conceptualization and qualitative understanding can be meaningfully made – an issue

---

[14]Generally, the agent's conceptualization of the world may be more coarse-grained than that of a well-informed observer. The agent might only distinguish between non-singleton equivalence classes of alternatives rather than between individual alternatives. In this case, only propositions expressible as unions of such equivalence classes would be conceptualized by the agent. According to Axiom 1 below, the agent would then be indifferent between alternatives in the same equivalence class.

which is partly philosophical and partly psychological. We flag it here as something that merits further investigation.

A third account draws on the concept of attentional salience as frequently used in psychology and behavioural economics. Among those propositions abstractly conceptualized by an agent and perhaps even 'understood' in some stronger, qualitative sense, only some are typically salient for the agent, in that the agent focuses on them or uses them as 'heuristics' or criteria in forming his or her preferences. Now the idea is that a proposition becomes motivating for an agent if and only if he or she focuses on it actively or uses it as a preference-formation heuristic or criterion. This account is consistent with the commonly made psychological assumption that the agent is boundedly rational, that is, unable or at least unlikely to give full and simultaneous attention to everything he or she conceptualizes or understands.[15] The account can also accommodate the important possibility that the agent actively engages in normative reflection on which propositions to take into account in forming his or her preferences. He or she might ask him- or herself which propositions are genuine normative reasons; we return to this possibility in our discussion of deliberation. Other accounts of the sources of motivation can be found in the philosophical literature. Whichever account we adopt, however, the basic idea that the agent's preferences depend on his or her set of motivating reasons is a very natural one.

4   The psychological states of an agent

From a third-person perspective, a full theory of an agent requires the ascription of an entire family of preference orders to that agent, consisting of one preference order $\succsim_M$ for each psychologically possible set of motivating reasons M. As we have noted, each such set corresponds to a particular motivational state of the agent, and in any such state, the agent holds only one preference order. What the reference to an entire family of preference orders captures is the idea that the agent may have a disposition to change his or her preferences in certain ways when his or her motivational state changes. The policy maker in our earlier example may prefer an invasion of an oil-producing country to an investment in renewable resources if he or she is motivated only by whether the policy supports current consumption levels of cheap oil. This preference may change, however, if he or she becomes motivated also by whether the policy leads to war. Of course, the agent him- or herself need not – and typically will not – be consciously aware of the entire family of preference orders ascribed to him or her by our theory. But from a theoretical perspective, we would like to account

---

[15] A prominent account of rational choice based on heuristics has been developed by Gerd Gigerenzer, Peter M. Todd and the ABC Group, *Simple heuristics that make us smart* (New York: Oxford University Press, 1999). See also Gerd Gigerenzer and Reinhard Selten (eds.), *Bounded rationality: The adaptive toolbox* (Cambridge, MA: MIT Press, 2001).

for the agent's preferences across variations in his or her motivational state.

In order to ascribe to the agent one preference order $\succsim_M$ for each possible set of motivating reasons M, we must specify what the possible such sets are. In other words, we must say something about what psychological states the agent can be in. In the simplest case, every subset of $\mathcal{P}$, the underlying set of possible reasons, constitutes a possible motivating set. But we have already noted that there may be psychological constraints on which propositions can motivate an agent, and under what conditions, and as a result, not every subset of $\mathcal{P}$ needs to be a possible specification of M. For instance, some propositions may never motivate the agent in conjunction with certain others. Some reasons may crowd out others, such as economic self-interest driven reasons versus charitable ones. Similarly, there may be propositions that can motivate the agent *only* in conjunction with certain others, and so on.

Thus, in the general case, the set of all possible sets of motivating reasons, which we call $\mathcal{M}$, may be smaller than the set of all subsets of $\mathcal{P}$. For expositional simplicity, we make a regularity assumption about the possible sets of motivating reasons:

Regularity assumption. The different possible sets of motivating reasons (that is, the elements of $\mathcal{M}$) form a lattice, that is:

  (i) if M_1 and M_2 are possible sets of motivating reasons, then so is $M_1 \cap M_2$ (that is, $\mathcal{M}$ is closed under intersection);

 (ii) if M_1 and M_2 are possible sets of motivating reasons, then so is $M_1 \cup M_2$ (that is, $\mathcal{M}$ is closed under union).

In the baseline case in which all subsets of $\mathcal{P}$ are possible sets of motivating reasons, this regularity assumption is trivially satisfied. In fact, our formal analysis requires only something weaker than this assumption. Theorem 1 below and its subsequent extension 1* use only part (i), and Theorem 2 and its extension 2* use only a weakened variant of part (ii).[16] Thus the 'crowding-out' or 'crowding-in' effects we have referred to can be captured by our analysis.[17]

5   Two axioms on reasons and preferences

We are now in a position to introduce our two central axioms on the relationship between an agent's set of motivating reasons and his or her preferences. The idea underlying both axioms is that an agent's preference for or against an alternative as compared with others is driven by the motivating reasons that are true of the alternative. The first axiom concerns the case in which the exact same motivating reasons are true of a given pair of alternatives.

---

[16]This weakened variant requires that if M_1 and M_2 are possible sets of motivating reasons, then so is *some* superset of $M_1 \cup M_2$. This is satisfied, e.g., if $\mathcal{P}$ is a possible motivating set.

[17]By suitably amending our axioms below, we can also say something about the case in which neither part of the assumption holds, but we do not discuss the details here.

Axiom 1. The agent is indifferent between any pair of alternatives of which the same motivating reasons are true. Formally, for any x and y in X and any M in $\mathcal{M}$, if $\{R \in M : R \text{ is true of } x\} = \{R \in M : R \text{ is true of } y\}$, then $x \sim_M y$.

Axiom 1 is true almost by definition under two of our three illustrative accounts of when a proposition becomes motivating for the agent. Consider the 'conceptualization' account. If the set of propositions motivating the agent coincides with the set of propositions he or she conceptualizes abstractly, this means that the agent does not distinguish between alternatives that differ at most with respect to propositions that are *not* motivating reasons – these propositions are not conceptualized – and hence indifference between such alternatives is natural.[18] Similarly, consider the 'attentional salience' account. Suppose the agent gives no attention to any propositions that do not motivate him or her or does not use them as preference-formation heuristics or criteria, either due to cognitive limitations or because he or she has explicitly set them aside. Then it is only natural that he or she will be indifferent between alternatives that differ at most with respect to such non-motivating propositions. Under the remaining illustrative account of the sources of motivation, the 'qualitative understanding' account, Axiom 1 becomes a substantive – and we think interesting – psychological hypothesis. On this account, Axiom 1 says that an agent's preferences between alternatives are fully determined by those properties of the alternatives the agent qualitatively understands, while any properties not understood in this manner make no difference.

Axiom 2 concerns the case in which the agent's set of motivating reasons in relation to the alternatives in X grows, but none of the newly added reasons is true of a given *pair* of alternatives x and y.

Axiom 2. If additional reasons become motivating for the agent, but none of them are true of a given pair of alternatives, then the agent's preference over that pair remains unchanged. Formally, for any x and y in X and any M and M' in $\mathcal{M}$ with $M' \supseteq M$, if no R in $M' \backslash M$ is true of x or y, then $x \succsim_M y \Leftrightarrow x \succsim_{M'} y$.

This implies, for example, that if the proposition 'Bordeaux wine is served at dinner' attains motivational relevance for the agent, this does not affect his or her preference between any two dinner plans not involving any wine. The axiom is plausible under each of our three illustrative accounts of the sources of motivation, especially in light of the idea that an agent's preferences between alternatives are driven by the motivating reasons that are true of those alternatives.

---

[18]Indifference follows strictly under the following conditions: (i) the agent distinguishes only between possibly non-singleton equivalence classes of alternatives, so that his or her preferences over individual alternatives are induced by preferences over these equivalence classes, and (ii) his or her motivating reasons are precisely the propositions expressible as unions of such equivalence classes.

Apparent counterexamples to Axiom 2, that is, preference changes seemingly driven by the addition of reasons not true of any of the alternatives in question, typically involve an under-specification of the reasons that are being added to the agent's motivating set. To illustrate, consider the following apparent counterexample. An agent prefers having dinner at McDonald's to dining at an organic vegan teetotalers' restaurant. But when he adopts the proposition 'Bordeaux wine is served at dinner' as a further motivating reason in deliberating about the alternatives, this prompts in him a more sophisticated attitude towards food and thereby reverses his preference between McDonald's and the organic alternative. Since neither restaurant is licenced to serve any wine, Axiom 2 appears to be violated by this preference change. This appearance, however, rests on an under-specification of the additional motivating reasons that lead to the agent's new psychological state. Implicit in the example is the thought that, along with the proposition 'Bordeaux wine is served at dinner', a second proposition such as 'the food is sophisticated' has also become motivationally relevant, and it is the latter reason that is responsible for the preference change. This is consistent with Axiom 2, since one of the two dinner options offers sophisticated food.

6 A general representation of reason-based preferences

What is the consequence of the two axioms we have introduced? Our first representation theorem shows that, if (and only if) an agent satisfies them, his or her preferences across all variations in motivating reasons can be parsimoniously represented in terms of a single binary relation, to be called a *weighing relation*, which ranks different possible combinations of reasons relative to one another. Formally, a *possible combination of reasons* is a consistent subset of $\mathcal{P}$, such as the set of propositions 'I have enough food' and 'I am healthy'. (A set of propositions is *consistent* if there is an alternative x in X of which all the propositions in it are true.)

Theorem 1. The agent's preference orders $\succsim_M$ across all variations in the set of motivating reasons M in $\mathcal{M}$ satisfy Axioms 1 and 2 *if and only if* there exists a weighing relation, denoted $\geq$, over all possible combinations of reasons such that, for each M in $\mathcal{M}$,

$$x \succsim_M y \iff \{R \in M : R \text{ is true of } x\} \geq \{R \in M : R \text{ is true of } y\} \text{ for all } x, y \text{ in } X.$$

Informally, the weighing relation whose existence is ensured by the two axioms delivers the agent's preferences as follows: in any motivational state, the agent prefers an alternative x to another alternative y if and only if the weighing relation ranks the combination of motivating reasons true of x above the combination of motivating reasons true of y. The two combinations of reasons that are being weighed relative

to each other can be interpreted as characterizing x and y through the lens of the agent's motivational state.

Before we turn to the interpretation of this relation, it is useful to give an example. Consider a simple case in which there are only four possible alternatives over which the agent has preferences:

Health care is available to everyone and cheap for me (ac).

Health care is not available to everyone but cheap for me ($\neg ac$).

Health care is available to everyone but not cheap for me ($a\neg c$).

Health care is neither available to everyone nor cheap for me ($\neg a\neg c$).

For simplicity, suppose, further, there are only two possible reasons in $\mathcal{P}$, namely

A : health care is available to everyone (formally $\{ac, a\neg c\}$), and

C : health care is cheap for me (formally $\{ac, \neg ac\}$),

but any set of them can be motivating, that is, $\mathcal{M}$ consists of all subsets of $\mathcal{P}$. Now imagine that the agent's preferences across variations in his or her motivating reasons are as follows:

$$M = \{A, C\} \quad \Rightarrow \quad ac \succ_M a\neg c \succ_M \neg ac \succ_M \neg a\neg c;$$
$$M = \{A\} \quad \Rightarrow \quad ac \sim_M a\neg c \succ_M \neg ac \sim_M \neg a\neg c;$$
$$M = \{C\} \quad \Rightarrow \quad ac \sim_M \neg ac \succ_M a\neg c \sim_M \neg a\neg c;$$
$$M = \varnothing \quad \Rightarrow \quad ac \sim_M a\neg c \sim_M \neg ac \sim_M \neg a\neg c.$$

One can verify that these preferences do indeed satisfy Axioms 1 and 2, so that our theorem applies. What, then, does the agent's weighing relation look like? It is easy to check that the agent's family of preference orders just displayed can be represented by a single weighing relation $\geq$ over possible reason combinations that satisfies

$$\{A, C\} > \{A\} > \{C\} > \varnothing,$$

where $>$ denotes the strict relation induced by $\geq$. Thus the reason combination $\{A, C\}$ is ranked first, the combination $\{A\}$ second, the combination $\{C\}$ third, and the empty combination last, which captures a particular way of weighing these possible combinations of reasons relative to each other.

But how can the agent's weighing relation be interpreted? We can distinguish between at least two broadly different kinds of interpretations. According to the first, which may be described as 'cognitivist', a weighing relation encodes a particular set of judgments about the relative 'goodness' of different possible reason combinations. Specifically, $S_1 \geq S_2$ is taken to mean that S_1 is a (weakly) better combination than S_2. Depending on the precise variant of this interpretation, the truth-conditions of

these judgments, if there are any, may be either agent-independent or agent-dependent facts about the goodness of different possible reason combinations. According to the second kind of interpretation, which we may call 'non-cognitivist', a weighing relation encodes the agent's dispositions to prefer certain combinations of reasons to others when the reasons contained in them are motivating. Here, $S_1 \geq S_2$ is taken to mean that the combination S_1 is (weakly) preferred to the combination S_2, assuming all reasons in S_1 and all those in S_2 are motivating. Again, different variants of this interpretation are conceivable, depending on what precisely is understood by a preference over possible reason combinations.

Regardless of the interpretation adopted, our theorem shows that when the relationship between an agent's set of motivating reasons and his or her preferences is governed by our two axioms, these preferences can be parsimoniously represented in terms of a single weighing relation whose relata are possible combinations of reasons. Furthermore, this weighing relation is essentially unique. (It is unique on the pairs of reason combinations needed to generate the agent's preference orders across variations in motivating reasons.[19]) In short, our theorem delivers a simple representation of what is by itself a rich structure, namely the agent's family of preference orders across all variations in his or her motivational state.

7 Is the weighing relation transitive?

Although we have considered different possible interpretations of the agent's weighing relation, we have not said anything yet about its formal properties. Most importantly, is it actually an order over all possible combinations of reasons? In other words, is it a complete and transitive binary relation? Completeness turns out to be not much of a problem since the weighing relation can always be defined so as to (weakly) rank all pairs of possible reason combinations. Surprisingly, however, the conditions introduced so far do not guarantee that the relation will always be transitive, despite the fact that all the actual preference orders generated by it are transitive. So how can an intransitivity in the weighing relation occur, and when is it ruled out?

To address these questions, it is helpful to begin with an example. Consider an agent who forms preferences over three types of cars available on the market:

a Monster Hummer, which is fast, big, but not environmentally friendly ($fb\neg e$);

a Sports Beetle, which is fast, not big, but environmentally friendly ($f\neg be$);

a Family Hybrid, which is not fast, but big and environmentally friendly ($\neg fbe$).

---

[19] These are all the pairs of reason combinations expressible as $\{R \in M : R$ is true of $x\}$ and $\{R \in M : R$ is true of $y\}$ for some x, y in X and M in $\mathcal{M}$. The weighing relation is underdetermined only with respect to those pairs of possible reason combinations that cannot be instantiated as true of some actual alternatives in X *and* simultaneously motivating, and such pairs do not really matter from the perspective of the agent's rational choices.

Thus, for the purposes of our example, any car available on the market has precisely two out of the three characteristics: fast (f), big (b), and environmentally friendly (e). Suppose, further, that a car's having any one of these characteristics can serve as a reason for or against preferring it, that is, the different possible reasons in $\mathcal{P}$ are

$$F \quad : \quad \text{the car is fast (formally } \{fb\neg e, f\neg be\}),$$
$$B \quad : \quad \text{the car is big (formally } \{fb\neg e, \neg fbe\}),$$
$$E \quad : \quad \text{the car is environmentally friendly (formally } \{f\neg be, \neg fbe\}).$$

Moreover, we assume that any set of these propositions can be motivating, that is, $\mathcal{M}$ contains all subsets of $\mathcal{P}$. Now it is entirely conceivable that the agent's family of preference orders across variations in motivating reasons is the following:

$$M = \{F, B, E\} \quad \Rightarrow \quad \text{Hummer} \sim_M \text{Beetle} \sim_M \text{Hybrid},$$
$$M = \{F, B\} \quad \Rightarrow \quad \text{Hummer} \succ_M \text{Beetle} \succ_M \text{Hybrid},$$
$$M = \{B, E\} \quad \Rightarrow \quad \text{Hybrid} \succ_M \text{Hummer} \succ_M \text{Beetle},$$
$$M = \{F, E\} \quad \Rightarrow \quad \text{Beetle} \succ_M \text{Hybrid} \succ_M \text{Hummer},$$
$$M = \{F\} \quad \Rightarrow \quad \text{Hummer} \sim_M \text{Beetle} \succ_M \text{Hybrid},$$
$$M = \{B\} \quad \Rightarrow \quad \text{Hummer} \sim_M \text{Hybrid} \succ_M \text{Beetle},$$
$$M = \{E\} \quad \Rightarrow \quad \text{Beetle} \sim_M \text{Hybrid} \succ_M \text{Hummer},$$
$$M = \varnothing \quad \Rightarrow \quad \text{Hummer} \sim_M \text{Beetle} \sim_M \text{Hybrid}.$$

One can check without much difficulty that these preferences satisfy Axioms 1 and 2,[20] and so, by Theorem 1, they are representable in terms of a single underlying weighing relation $\geq$ over possible reason combinations. But what is this weighing relation? In order to be able to generate the agent's preferences just displayed, it must have all of the following properties:

$$\{F, B\} \equiv \{B, E\} \equiv \{F, E\},$$

$$\{F, B\} > \{F\} > \{B\},$$
$$\{B, E\} > \{B\} > \{E\},$$
$$\{F, E\} > \{E\} > \{F\},$$

$$\{F\} > \varnothing,$$
$$\{B\} > \varnothing,$$
$$\{E\} > \varnothing.$$

Here $\equiv$ denotes the symmetrical relation induced by $\geq$, and $>$ denotes the strict relation, as before. From the second, third and fourth lines, it follows immediately

---

[20] This is straightforward in the case of Axiom 1. To see that Axiom 2 is satisfied, note that the structure of the example implies that if x, y in X, M in $\mathcal{M}$ and R in $\mathcal{P}\backslash M$ are such that R is true of neither x nor y, then x and y must be identical ($x = y$), so that $x \sim_M y$ and $x \sim_{M \cup \{R\}} y$.

that the weighing relation violates transitivity, since $\{F\} > \{B\}$, $\{B\} > \{E\}$, and yet $\{E\} > \{F\}$. Of course, an intransitive weighing relation is harder to interpret than a transitive one, particularly if one chooses to adopt a cognitivist interpretation. The lesson to learn, however, is that the conditions used in our first representation theorem are simply not enough to rule out an intransitive weighing relation, though they are fully compatible with an agent's having a transitive such relation.[21]

What is the source of the intransitivity in the agent's weighing relation in our example? Imagine that, contrary to the assumptions made, there existed additional cars of which precisely one or none of the three possible reasons is true: one car that is only fast (a Ferrari), one that is only big (an Old Diesel Van), one that is only environmentally friendly (an Eco-Prius), and one without any of these properties (an East German Trabant). Then the agent's preference order over the different cars when motivated by all three reasons would constrain his or her weighing relation to rank all possible combinations of reasons, including the three singletons $\{F\}$, $\{B\}$ and $\{E\}$, transitively. The intransitivity identified in our example would disappear. The counterfactual stipulation just made would give the agent a kind of 'Olympian perspective' from which he or she would be able to consider one alternative corresponding to each possible combination of reasons, which instantiates all and only the reasons in it, and thereby to rank all these sets transitively. Generalizing from this observation, we can conjecture that an intransitivity in the agent's weighing relation can occur precisely if this Olympian perspective is not available.

Our second representation theorem confirms this conjecture. Call the set $\mathcal{P}$ of possible reasons *weakly independent* if every consistent subset S of $\mathcal{P}$, representing a possible reason combination, can be *instantiated precisely*: there is an alternative x in X of which, among the possible reasons in $\mathcal{P}$, precisely those in S, and no others, are true. To illustrate, weak independence is violated in our example of the three cars: although each of the singleton sets $\{F\}$, $\{B\}$ and $\{E\}$ (as well as the empty set) is consistent, thereby representing a possible combination of reasons, no cars instantiate any of these combinations precisely. By contrast, in the augmented example in which cars instantiating them are stipulated to exist, weak independence is satisfied.

Theorem 2. Suppose the set of possible reasons $\mathcal{P}$ is weakly independent. Then the agent's preference orders $\succsim_M$ across all variations in the set of motivating reasons M in $\mathcal{M}$ satisfy Axioms 1 and 2 *if and only if* there exists a *weighing order* (a complete

---

[21] An intransitivity in the weighing relation does *not* give rise to an intransitivity in the agent's preferences, provided these are defined relative to a fixed set of motivating reasons. If the consideration of different pairs of alternatives changed the set of motivating reasons, an intransitivity might surface. This would happen, e.g., if the comparison of $f\neg be$ and $\neg fbe$ made the set $\{F,B\}$ motivating (leading to a preference for $f\neg be$ over $\neg fbe$), the comparison of $\neg fbe$ and $fb\neg e$ made $\{F,E\}$ motivating (leading to a preference for $\neg fbe$ over $fb\neg e$), and the comparison of $fb\neg e$ and $f\neg be$ made $\{B,E\}$ motivating (leading to a preference for $fb\neg e$ over $f\neg be$). We return to this issue in Section 12.

and transitive weighing relation), denoted $\geq$, over all possible combinations of reasons such that, for each M in $\mathcal{M}$,

$$x \succsim_M y \;\Leftrightarrow\; \{R \in M : R \text{ is true of } x\} \geq \{R \in M : R \text{ is true of } y\} \text{ for all } x,y \text{ in } X.$$

This theorem confirms that in our example of the three cars the lack of weak independence of the set of possible reasons is indeed to blame for the unavailability of the Olympian perspective needed to ensure a transitive weighing relation. Conversely, the satisfaction of weak independence, as in the augmented car example, is enough to guarantee the transitivity of the weighing relation. As in our earlier representation theorem, the weighing relation – now a weighing order – is essentially unique.

8 Additive weighing of reasons

It is worth drawing attention to an important special case of an agent who is rational in the sense of our theory, and to whom our representation results therefore apply. Consider an agent whose reason-based preference formation works as follows. The agent implicitly assigns a particular numerical weight to each of the possible reasons in $\mathcal{P}$. Some possible reasons get assigned a positive weight, others a negative one. For example, the proposition 'there is peace' will presumably have a positive weight (counting in favour of any alternative of which it is true), each of the propositions 'there is not enough food available' and 'I am hungry' a negative one (counting against any alternative of which it is true). Now the agent prefers an alternative to another just in case the sum-total of the weights of the reasons true of the first alternative and motivating for him or her exceeds the same sum-total for the second. In each case, the sum-total encompasses reasons with a positive weight as well as reasons with a negative one.

To describe this process formally, we introduce a function w which assigns to each possible reason R in $\mathcal{P}$ a real number $w(R)$, interpreted as the weight of R. Of course, this function may differ from agent to agent. For each set of motivating reasons M in $\mathcal{M}$, the agent's preference order $\succsim_M$ is now given as follows:

$$x \succsim_M y \Leftrightarrow \sum_{R \in M : R \text{ is true of } x} w(R) \geq \sum_{R \in M : R \text{ is true of } y} w(R) \quad \text{for all } x,y \text{ in } X.$$

Our two axioms are clearly satisfied here, and the agent's weighing relation over possible combinations of reasons can easily be derived from the weights assigned to each of the different reasons contained in them. Specifically, one such combination is ranked over another by the agent's weighing relation if and only if the sum-total of weights assigned to the reasons in the first combination exceeds that for the second, or formally:

$$S_1 \geq S_2 \Leftrightarrow \sum_{R \in S_1} v(R) \geq \sum_{R \in S_2} v(R) \quad \text{for any consistent } S_1, S_2 \subseteq \mathcal{R}.$$

To illustrate, recall our earlier example of the agent holding preferences over the four alternatives corresponding to the different possible truth-value combinations of the propositions 'health care is available to everyone' and 'health care is cheap for me'. In that example, the agent's preferences and underlying weighing relation can be represented in terms of the assignment of suitable weights to individual reasons. We obtain a correct representation of the given preferences, for instance, by assigning a weight of 2 to the reason 'health care is available to everyone' (A) and a weight of 1 to the reason 'health care is cheap for me' (C).

To be sure, an agent whose reason-based preference formation works like this is only a special case of an agent as described by our theory, since, in the current example, reasons have an 'additive separability' property they need not have in general: in this case of separability, the weight the agent assigns to any motivating reason is independent of what other reasons are motivating for him or her. In the general case permitted by our theory, there is no such restriction.

Nonetheless, the separable case we have flagged is important since the additive balancing of reasons that goes on here captures what in philosophical discussions is often described as the weighing of *pro tanto* reasons for and against some object of choice.[22] Indeed, it is only in the context of separability that any given reason can unambiguously be said to 'count in favour of' or 'against' the alternatives of which it is true. Without separability, the question of whether a reason counts for or against those alternatives depends also on which other reasons are present. The denial of separability is often described as 'particularism' or 'holism' about reasons, its affirmation as 'generalism' or 'atomism'.[23]

9   Explanation versus justification

It is appropriate at this point to revisit the interpretation of our reason-based approach to the theory of rational choice. Although we have distinguished between motivating and normative reasons, we have focused on the former, interpreting our framework as capturing the relationship between an agent's motivating reasons and his or her actual preferences. This focus makes sense from a positive, social-scientific perspective, from which we are primarily interested in *explaining* why agents make the choices they make, and sometimes in predicting those choices. From a normative perspective, however, we would also like to assess whether these or any other choices are *justified* – or at least whether they are *justifiable* – and whether the agent has

---

[22] See, e.g., the discussion of *pro tanto* reasons in Broome, 'Reasons' (op. cit.).

[23] For an excellent formal analysis of these views, see Campbell Brown, 'The Composition of Reasons' (University of Edinburgh, 2009). See also Ridge, 'Reasons for Action: Agent-Neutral vs. Agent-Relative' (op. cit.). Although particularism/holism and generalism/atomism are usually defined as views about *normative* reasons, analogous definitions can be given in the context of *motivating* reasons.

made them for the right reasons. To address these questions, we need to say more about normative reasons. While reason-based *explanations* must refer to motivating reasons, reason-based *justifications* require a reference to normative reasons.[24]

As already mentioned but not yet developed, our formal framework can be reinterpreted to address the relationship between an agent's normative reasons and the preferences he or she *ought* to have, rather than the relationship between an agent's motivating reasons and his or her actual preferences, on which we have focused so far. To sketch that reinterpretation, we must read the symbol $\succsim$ as representing the preference order the agent *ought* to have – which we may call his or her *ideal* preference order – rather than the one he or she *actually* has. And to indicate that this ideal preference order depends on the agent's *normative* reasons, we must now append a subscript N to $\succsim$, interpreting $\succsim_N$ as the preference order the agent ought to have in the event that N is his or her set of normative reasons in relation to the alternatives in X. Different accounts of what qualifies as a normative reason for an agent, how the 'ought' is to be understood (whether as a rational or a moral one, for instance), and how the set N must be specified, correspond to different variants of this interpretation.[25] Once the interpretation is settled, however, our formal framework can be taken to describe how the ideal preference order $\succsim_N$ varies with variations in the set of normative reasons N.

Just as the agent's actual preference order can be represented as being determined by the agent's motivating reasons, together with his or her underlying weighing relation, so the ideal preference order can now be represented as being determined by the relevant normative reasons, again on the basis of *some* underlying weighing relation. Of course, this formal analysis does not settle the question of what the right normative reasons are in any given context, or what weighing relation to use for ranking different combinations of reasons relative to one another. Different normative background theories will give different answers to these questions. Our theory only provides a formal calculus for linking normative reasons with ideal preferences, based on some underlying weighing relation. But the theory can be supplemented with normative constraints on the weighing relation, and with criteria for identifying the right set – or the permissible sets – of normative reasons in any given context.

Aided by this dual interpretability of our formal framework, we can return to the distinction between reason-based explanation and reason-based justification. To *explain* an agent's preferences, we simply need to show that these preferences are determined by the agent's motivating reasons, on the basis of his or her underlying

---

[24] On the distinction between reason-based explanation and reason-based justification, see also Lenman, 'Reasons for Action: Justification vs. Explanation' (op. cit.).

[25] On some accounts, N depends, at least partly, on what is accessible to the agent, so that, e.g., considerations not knowable by the agent cannot be normative reasons for him or her; on others, N is independent of this question.

weighing relation. We know from our representation theorems that, as soon as Axioms 1 and 2 are satisfied, such a reason-based explanation can always be given. However, whether these preferences – for example, the preference for x over y – are also *justified* depends on whether the agent's actual set of motivating reasons is also the right set of *normative* reasons – or at least a permissible such set – and whether the agent's weighing relation obeys the relevant normative constraints on how to weigh different combinations of reasons relative to one another. When the agent's actual preferences are justified, they coincide with his or her ideal preferences, that is, with the preferences he or she ought to have.

Recall our example of a policy maker who is deciding which foreign policy to support. As we have noted, he or she may support the invasion of an oil-producing country because this promises to make cheap oil available. But since it is hard to think of any mainstream theory of just war that would deem an invasion permissible on those grounds, the present case illustrates how reason-based explanation and reason-based justification can come apart. While the prospective availability of cheap oil constitutes a motivating reason *explaining* the policy maker's preference, it is by no means a normative reason *justifying* it.

We are also able to describe cases in which an agent's preferences are at least *justifiable*, even when the more stringent conditions of actual justification are not met. An agent's preferences are *justifiable* if there *exist* a right or permissible set of normative reasons and an acceptable weighing relation that would give rise to those preferences – more precisely, that would entail ideal preferences that coincide with those actual preferences – even if this is not the actual way the agent has arrived at them. People often exploit this kind of justifiability when they respond to criticism of their choices or actions by '*ex post* rationalization', that is, by pretending to have been motivated by normative reasons that would have justified their choices when these were not the actual motivating reasons.

In sum, our theory not only allows us to distinguish between explicable and justifiable choices, and between choices made for the right and the wrong reasons, but it also gives expression to the less commonly recognized possibility that an agent is motivated by the right normative reasons but governed by the wrong underlying weighing relation, or that the agent's being motivated by the wrong reasons – motivating reasons that are not genuine normative reasons in the given context – goes along with his or her having the right weighing relation. While the distinctions between reason-based explanation and reason-based justification and between acting for the right and the wrong reasons are familiar from the existing philosophical literature, the role played by the agent's weighing relation, and the additional complexities that open up once we subject this to a normative assessment as well, are not made explicit in the existing literature. It should therefore be evident that our proposed theory offers useful conceptual resources for addressing those under-researched issues.

10 Deliberation and disagreements

As our theory allows us to distinguish between reason-based explanation and reason-based justification, it can also shed light on the role played by reasons in an agent's rational deliberation about his or her preferences and on how reasons might be relevant to resolving disagreements between different agents' preferences. Standard rational choice theory, as we have noted, takes preferences over possible worlds or fully specified outcomes – the *alternatives* in our theory – to be fundamental and unchangeable and thus cannot explain how deliberation, either within an individual or in a group, could ever lead to any revisions of those preferences. Changes in preferences, on this assumption, are only possible at the level of derived preferences over uncertain prospects and must stem from changes in the agent's beliefs about which outcomes are likely to result from those prospects. By implication, when individuals engaged in collective deliberation have different preferences, all they can do is to resolve any informational differences between them and, if this does not help to reach agreement (because the disagreement was not due to different information), to aggregate their conflicting fundamental preferences into overall collective preferences. This, however, does not resolve the individual-level disagreement; at best, it generates some collective compromise. Furthermore, the process is notoriously vulnerable to the paradoxes and impossibility results of aggregation familiar from social choice theory in the tradition of Condorcet and Arrow.[26]

This standard picture fails to account for the possibility that an agent's preferences may change as a result of changes in his or her motivating reasons in relation to the given alternatives. Such changes may, in turn, be prompted by various experiences and especially by individual or collective deliberation. We can think, for example, of a capitalist businessman who, after surviving a plane crash, consciously forms a preference for a life devoted to charity over a life driven by income maximization, or a workaholic who, after recovering from an illness, consciously abandons his or her work-oriented preferences.[27] Similarly, group deliberation may change participants' assessment of fundamental alternatives, for instance by making previously overlooked aspects of those alternatives salient to them.[28] Arguably, these agents have not merely

---

[26] Kenneth Arrow, *Social Choice and Individual Values* (New York: Wiley, 1951/1963).

[27] For a more detailed discussion, see Franz Dietrich and Christian List, 'A Model of Non-Informational Preference Change', *Journal of Theoretical Politics* (forthcoming).

[28] For related discussions, see David Miller, 'Deliberative Democracy and Social Choice', *Political Studies* 40 (special issue) (1992), pp. 54-67; Jack Knight and James Johnson, 'Aggregation and Deliberation: On the Possibility of Democratic Legitimacy', *Political Theory* 22 (1994), pp. 277-296; John S. Dryzek and Christian List, 'Social Choice Theory and Deliberative Democracy: A Reconciliation', *British Journal of Political Science* 33 (1) (2003), pp. 1-28; and Christian List, Robert Luskin, James Fishkin and Iain McLean, 'Deliberation, Single-Peakedness, and the Possibility of Meaningful Democracy: Evidence from Deliberative Polls', working paper, London School of Economics and Stanford Center for Deliberative Democracy (2000/2007).

learnt new information – although some of their beliefs may have changed along the way – but new reasons, such as other-regarding or non-economic reasons, have become motivating for them.

Our reason-based theory of rational choice allows us to capture these phenomena. It allows us to distinguish between information-based and reason-based deliberation, and to acknowledge mixtures of the two. Information-based deliberation, as recognized by standard rational choice theory, takes place whenever an agent rationally revises his or her beliefs in response to new information or evidence, which may affect the agent's derived preferences over uncertain prospects, but not his or her preferences over fundamental alternatives. Reason-based deliberation, on the other hand, takes place when the agent rationally revises his or her preferences at the fundamental level in response to changes in his or her set of motivating reasons.[29]

On this richer picture, an agent can enter into a conscious deliberative process with the aim of identifying which propositions to take into account in forming his or her preferences over a given set of alternatives. Although someone not engaged in explicit deliberation may sometimes find him- or herself simply motivated by some reasons rather than others, we need not assume that one's motivating reasons are always outside one's control. An agent can deliberately interrogate him- or herself about which propositions are genuine normative reasons for him or her in relation to some alternatives, and thereby exercise some influence over which reasons come to motivate him or her.

The possibility of changing one's motivating reasons through deliberation is consistent with each of our three illustrative accounts of when a proposition becomes motivating, although the 'attentional salience' account arguably captures the role of normative reflection best. Let us begin with the 'conceptualization' account, according to which a proposition attains motivational relevance for an agent as soon as he or she conceptualizes it abstractly. In line with this account, deliberation may affect an agent's set of motivating reasons by refining his or her conceptual abilities, that is, by helping him or her to distinguish between alternatives of which certain previously unconceptualized propositions are true and alternatives of which they are not. But since conceptualizing a proposition is not the same as judging that it matters normatively, this account does not capture how normative reflection in particular can affect an agent's motivating reasons.

Next, consider the 'qualitative understanding' account, according to which a proposition becomes motivating for an agent when he or she qualitatively – and not just abstractly – understands it. Although developing this idea would require further elaboration, it is plausible that deliberation, at least when construed broadly, is not restricted to the exchange of information or to abstract conceptual reasoning, but

---

[29] For an earlier taxonomy of informational, argumentative, reflective, and social aspects of deliberation, see Dryzek and List, 'Social Choice Theory and Deliberative Democracy' (op. cit.).

that it can also make an agent imagine various scenarios vividly and thereby enhance his or her qualitative understanding of some of the propositions true of those scenarios.[30] Think, for example, of how a startling personal report of someone's experience – say, the experience of war – can evoke in the listener a qualitative sense of what it would have been like to go through that experience oneself. Like abstract conceptualization, however, having a qualitative understanding of something is distinct from judging that it matters normatively, and so this account equally fails to capture the role of normative reflection in shaping an agent's set of motivating reasons.

Finally, in the case of the 'attentional salience' account of motivation, it should be evident that deliberation can affect an agent's set of motivating reasons. If propositions motivate an agent when they are salient in the right way, then any activity, such as deliberation, that involves giving careful attention to various propositions can confer the required salience on some of them and de-emphasize others. Moreover, the variant of this account that ascribes to the agent active control over which propositions to use in forming his or her preferences can also capture the idea that normative reflection may affect the agent's set of motivating reasons. It is an open question whether, and how, deliberation may affect not only the agent's motivating reasons, as argued, but also his or her underlying weighing relation. We need not take a view on this issue here, except to mention it for further investigation.

The present observations suggest a more nuanced perspective on disagreements between different agents' preferences. As we have seen, when agents disagree in their preferences even after exchanging all relevant information, standard rational choice theory offers no further resources for resolving that disagreement – except to apply some method of 'brute' aggregation for arriving at some overall collective compromise. Our theory, by contrast, allows us to identify whether the disagreement stems from differences in the agents' sets of motivating reasons – perhaps along with differences in their judgments on which propositions are genuine normative reasons – or from differences in their underlying weighing relations, or both.

If it stems from differences at the level of reasons, it falls under the scope of deliberation in the broadened sense of our theory. The agents can then deliberate about which propositions are genuine normative reasons that should be used in forming their preferences over the relevant alternatives, and if they reach agreement on this matter, their original disagreement will have been resolved. But even if they cannot agree on the right normative reasons, their disagreement will have been made more tractable. Its source will have been identified, which means that it no longer needs to be attributed to a brute difference in tastes. Much of the recent debate on the idea of 'public reason' can be understood in these terms: the aim is to come up with

criteria for determining which reasons for and against the alternatives are publicly acceptable – that is, which reasons can be invoked to justify one's preferences in public deliberation – and which are not.

If the agents' disagreement stems from differences in their underlying weighing relations, on the other hand, the situation is more complicated. We have left it open to what extent deliberation can affect one's weighing relation, but it should be noted that many debates in moral philosophy can be understood as subjecting that relation to normative assessment as well. Such assessment takes place whenever the relative importance or weight of different possible reasons is being discussed.

Drawing on these observations, our theory further allows us to suggest a new approach to the aggregation of preferences: reason-based preference aggregation. Here each agent's preference order would be treated not as a fundamental and unchangeable input to the aggregation, but as being derived from two more fundamental inputs: a set of motivating reasons – or perhaps some judgments on what the right normative reasons are – and an underlying weighing relation. By making explicit and disentangling these two determinants behind any preference order, the informational basis for the aggregation would be enriched, which might allow us to find more compelling methods of aggregation. Although this proposal needs to be elaborated further, many theorists of 'public reason' may be attracted to the idea of aggregating preferences in a way that is sensitive to the reasons behind those preferences.

11   Preferences over uncertain prospects

A satisfactory theory of rational choice must say something not only about an agent's preferences over fundamental alternatives – possible worlds, states, or fully specified outcomes – but also about his or her preferences over uncertain prospects. An agent's objects of choice are often different possible actions, which correspond to different prospects. Each action usually has several possible outcomes, and the agent is at most able to assign probabilities to them. These probabilities normally represent the agent's beliefs about the likelihood of those outcomes, but they could also have an objective interpretation. In this section, we show how our theory can be extended so as to capture preferences over prospects in full generality. Less technically inclined readers may skip this section without losing the overall thread of our argument.

Formally, a *prospect* is a probability distribution over the alternatives in X, that is, a function P from the set X of alternatives into the interval $[0, 1]$ whose sum-total across all alternatives is 1.[31] We write $\mathcal{X}$ to denote the set of all prospects. We now assume that an agent's preference order $\succsim_M$, for any set of motivating reasons M, is defined not just over the alternatives in X, but over all prospects in $\mathcal{X}$. Moreover, we assume that, for each set of motivating reasons M, the agent's preferences are

---

[31] For simplicity, we require P to have finite support, i.e., $P(x) > 0$ for finitely many x in X.

'classical', in the sense that $\succsim_M$ ranks prospects according to the expectation of some 'utility' function from the set X of alternatives into the real numbers.

Again, we impose two axioms on the agent's preferences. The first is identical in meaning to our earlier Axiom 1. This is because it quantifies over all sure prospects, where a *sure prospect* is one that assigns probability 1 to a single alternative x in X and probability 0 to all others; it can thus be identified with that alternative x.

Axiom 1*. The agent is indifferent between any pair of sure prospects of which the same motivating reasons are true. Formally, for any sure prospects x and y in $\mathcal{X}$ and any M in $\mathcal{M}$, if $\{R \in M : R \text{ is true of } x\} = \{R \in M : R \text{ is true of } y\}$, then $x \sim_M y$.

The second axiom quantifies over *all* prospects, not merely the sure ones, but otherwise matches our earlier Axiom 2. To state this axiom, we define the probability that a prospect assigns to a given proposition as the sum of the probabilities it assigns to the alternatives of which that proposition is true.[32]

Axiom 2*. If additional reasons become motivating for the agent, but all of them are assigned zero probability by a given pair of prospects, then the agent's preference over that pair remains unchanged. Formally, for any prospects P and Q in $\mathcal{X}$ and any M and M' in $\mathcal{M}$ with $M' \supseteq M$, if all R in $M' \backslash M$ receive zero probability under P and Q, then $P \succsim_M Q \Leftrightarrow P \succsim_{M'} Q$.

When the agent's preferences satisfy Axioms 1* and 2*, two representation theorems hold, which are direct analogues of our earlier two theorems. Let us begin with the analogue of Theorem 1, which provides a representation of the agent's preferences in terms of a single weighing relation, this time defined not over possible combinations of reasons themselves, but more generally over probability distributions over such combinations.[33] Recall that, in Theorem 1, the preference between two alternatives x and y was determined by comparing the two combinations of reasons $\{R \in M : R \text{ is true of } x\}$ and $\{R \in M : R \text{ is true of } y\}$, characterizing the two alternatives from the perspective of the agent's motivational state. In the extension to the case of uncertainty, the preference between two prospects P and Q is determined by comparing two induced probability distributions over combinations of reasons, which we call P_M and Q_M. These can be seen as the probabilistic generalizations of the two combinations of reasons compared in Theorem 1, this time characterizing the prospects P and Q, rather than the alternatives x and y, from the perspective of the agent's motivational state. Specifically, P_M and Q_M assign to each possible combination of

---

[32]Formally, for any R in $\mathcal{P}$, $P(R) := \sum\limits_{x \in X : R \text{ is true of } x} P(x)$.

[33]A probability distribution over possible combinations of reasons is a function from the set of all possible combinations of reasons to the interval $[0, 1]$ which sums to 1, where, as before, only finitely many combinations of reasons are assigned non-zero probability.

reasons S the total probability (according to P and Q, respectively) of those alternatives x in X for which the propositions in S are precisely the agent's true motivating reasons.[34] In the special case in which P and Q are sure prospects – and thus identifiable with some alternatives x and y – the induced distributions P_M and Q_M assign probability 1 to $\{R \in M : R \text{ is true of } x\}$ and $\{R \in M : R \text{ is true of } y\}$, respectively. Now the analogue of Theorem 1 can be stated as follows:

Theorem 1*. The agent's preference orders $\succsim_M$ across all variations in the set of motivating reasons M in $\mathcal{M}$ satisfy Axioms 1* and 2* *if and only if* there exists a weighing relation, denoted $\geq$, over all probability distributions over possible combinations of reasons such that, for each M in $\mathcal{M}$,

$$P \succsim_M Q \;\Leftrightarrow\; P_M \geq Q_M \text{ for all } P, Q \text{ in } \mathcal{X},$$

where P_M and Q_M are the induced probability distributions just defined.

While Theorem 1* shows that the satisfaction of Axioms 1* and 2* is enough to render the agent's preferences representable in terms of a single weighing relation, we obtain a stronger representation when the set of possible reasons $\mathcal{P}$ is weakly independent, as before. Recall that weak independence of $\mathcal{P}$ means that, for every consistent subset S of $\mathcal{P}$, representing a possible reason combination, there is an alternative x in X of which, among the possible reasons in $\mathcal{P}$, precisely those in S are true. The analogue of Theorem 2 now yields a representation of the agent's preferences in terms of a *weighing function*, denoted W, over possible combinations of reasons, not just in terms of a weighing order $\geq$ over them.[35] Although such a weighing function W induces a weighing order $\geq$, which ranks probability distributions over possible reason combinations according to their expected weight under W, it contains more information than that order.

Theorem 2*. Suppose the set of possible reasons $\mathcal{P}$ is weakly independent. Then the agent's preference orders $\succsim_M$ across all variations in the set of motivating reasons M in $\mathcal{M}$ satisfy Axioms 1* and 2* *if and only if* there exists a real-valued weighing function, denoted W, over all possible combinations of reasons such that, for each M in $\mathcal{M}$, $\succsim_M$ ranks prospects in $\mathcal{X}$ according to the expected weight (according to W) of the combination of true motivating reasons.[36]

---

[34] P_M and Q_M are the projections of P and Q under the function that maps each alternative x in X to the combination of reasons $\{R \in M : R \text{ is true of } x\}$. For each possible combination of reasons S,

$$P_M(S) = \sum_{\substack{x \in X: \\ \{R \in M : R \text{ is true of } x\} = S}} P(x) \quad \text{and} \quad Q_M(S) = \sum_{\substack{x \in X: \\ \{R \in M : R \text{ is true of } x\} = S}} Q(x).$$

[35] The weighing function W over possible *combinations* of reasons should not be confused with the function w assigning weights to *individual* reasons in the additively separable case.

[36] This expected weight is the expectation of the induced utility function u_M mapping each al-

Thus the extension of our theory to the case of preferences over general prospects adds some further structure to the second representation theorem as compared with its earlier counterpart. In Theorem 2 above, as noted, weak independence of the set of possible reasons $\mathcal{P}$ ensured an *ordinal* representation of the agent's preferences in terms of a single weighing order over possible combinations of reasons. By contrast, Theorem 2* yields a *cardinal* representation of those preferences in terms of a single weighing function over possible combinations of reasons. The agent's preferences over prospects are then determined by the expected weight of the combination of true motivating reasons under those prospects.

We can redescribe this representation in another way. For each possible set of motivating reasons M, the weighing function W can be interpreted to induce a utility function u_M on the set of alternatives X. This utility function assigns to each alternative the weight (according to W) of the combination of reasons that are true of that alternative and motivating, formally

$$u_M(x) = W(\{R \in M : R \text{ is true of } x\}) \quad \text{for each } x \text{ in } X.$$

The agent's preference order $\succsim_M$ over prospects is then determined by the expectation of that utility function for the given prospects, formally

$$P \succsim_M Q \Leftrightarrow \sum_{x \in X} P(x) u_M(x) \geq \sum_{x \in X} Q(x) u_M(x) \quad \text{for all } P, Q \text{ in } \mathcal{X}.$$

The present results demonstrate not only that our theory can represent preferences over uncertain prospects as much as it can represent preferences over sure ones, but also that it properly generalizes standard rational choice theory. Indeed, standard rational choice theory emerges as the special case of our theory in which the set M of motivating reasons is assumed to be fixed and sufficiently large to impose no restrictions on the assignment of utilities to individual alternatives.

12   From preferences to choices

Up to now, we have focused on the relationship between reasons and preferences and left implicit how these relate to choices. To complete our theory, we need to address this final step. We concentrate once again on the positive interpretation of our theory, considering the path from motivating reasons, via actual preferences, to resulting choices, although our analysis can also be reinterpreted in normative terms, so as to capture the path from normative reasons, via ideal preferences, to the choices the agent ought to make.

---

ternative x in X to $W(\{R \in M : R \text{ is true of } x\})$. The weighing function W is unique up to positive affine transformations on the subdomain of those possible combinations of reasons needed to generate the agent's preferences. Each of the functions u_M (for all M in $\mathcal{M}$) is unique up to the same transformations on its entire domain.

The central observation is that the way decision theory formally relates preferences to choices carries over to reason-based preferences. Any preference order, including a reason-based one, induces a corresponding choice function. While a preference order represents certain *intentional attitudes* towards a given set of alternatives, a choice function encodes certain *choice dispositions* in relation to these alternatives. It specifies which alternative, or alternatives, would (or on a normative interpretation, should) be chosen from any available subset of the alternatives. Formally, a *choice function*, denoted C, assigns to each non-empty subset Y of X (of available alternatives) a set of one or more chosen alternatives from the ones in Y.

Suppose, for example, an agent is faced with a choice between different fruits, such as apples, bananas, and oranges. The agent's choice function represents which fruit(s) he or she would pick from any particular set of available ones, say from a particular fruit basket he or she is presented with. The function might look like this:

$$
\begin{aligned}
C(\{\text{apple,banana,orange}\}) &= \{\text{apple}\}; \\
C(\{\text{apple,banana}\}) &= \{\text{apple}\}; \\
C(\{\text{banana,orange}\}) &= \{\text{banana}\}; \\
C(\{\text{apple,orange}\}) &= \{\text{apple}\}; \\
C(\{\text{apple}\}) &= \{\text{apple}\}; \\
C(\{\text{banana}\}) &= \{\text{banana}\}; \\
C(\{\text{orange}\}) &= \{\text{orange}\}.
\end{aligned}
$$

That is, if all three fruits are available, the agent will choose an apple; if an apple and a banana are available, he or she will also choose an apple; if a banana and an orange are available, he or she will choose a banana; and so on. The present choice function is easily explicable: the agent simply prefers apples to bananas to oranges and always picks whichever fruit among the available ones is highest on this ranking.

Generally, a preference order $\succsim$ induces a choice function C as follows:[37]

$C(Y) = \{y \in Y : y \succsim x \text{ for all } x \in Y\}$ for any non-empty subset Y of X.[38]

Crucially, by saying that a preference order 'induces' a choice function, we do not settle the difficult philosophical question of whether the relationship between preferences and choices is best understood as causal or explanatory. Our formal analysis is intended to be neutral on this matter.

---

[37] Conversely, any choice function that represents sufficiently regular choice dispositions can be explained in terms of an underlying preference order or other binary relation. For a classic exposition of the relevant conditions, see Amartya Sen, 'Choice Functions and Revealed Preference', *Review of Economic Studies* 38 (3) (1971), pp. 307-317.

[38] To be well-defined, $C(Y)$ must always be non-empty. This requires that, for each subset Y of X, there is some alternative y in X that is maximal with respect to the preference order $\succsim$. This condition is trivially met if X is finite. It is also met if $\succsim = \succsim_M$, as defined in this paper, *and* M is a finite set of reasons, a plausible psychological hypothesis.

We can now ascribe to any agent modelled by our theory not only a family of preference orders $\succsim_M$ across the different possible sets of motivating reasons M in $\mathcal{M}$, but also a family of corresponding choice functions C_M across all M in $\mathcal{M}$. Each choice function C_M represents the agent's choice dispositions in the event that M is his or her set of motivating reasons in relation to the alternatives in X. (On the normative reinterpretation, a choice function C_N – notice the subscript N – would represent the choice dispositions the agent *ought* to have in the event that N is his or her set of normative reasons in relation to those alternatives.)

The resulting picture of rational choice should be clear: at any time, the agent is in a particular psychological state, represented by his or her set of motivating reasons in relation to the given alternatives, which, jointly with the agent's weighing relation, determines his or her preference order. This preference order then induces a choice function, which encodes how the agent would choose from any concrete set of alternatives. By implication, a change in the agent's set of motivating reasons can bring about not only a change in his or her preference order, but also a change in the choice function and thus in the resulting choice dispositions. In this way, motivating reasons can be viewed as motivating for preferences as well as for choices. (Similarly, on the normative reinterpretation, the set of normative reasons constrains the choices the agent *ought* to make, via determining his or her *ideal* preferences. Thus normative reasons, according to our theory, can be interpreted as reasons for preferences as well as for choices.) Using the technical tools from the previous section, this picture can be further extended to choices under uncertainty as well, but for simplicity we set these technicalities aside.

The picture just sketched, however, involves a simplifying assumption. By defining the agent's choice function C_M on the basis of the preference order $\succsim_M$, we have implicitly assumed that the set of motivating reasons M is given independently of the particular set of available alternatives to which the choice function is applied; it is defined simply in relation to the set X of all alternatives, while the set of available alternatives can be any subset Y of X. (In the normative case, a similar assumption is implicit in defining an ideal choice function C_N on the basis of an ideal preference order $\succsim_N$ for a set of normative reasons N that does not depend on the particular set Y of available alternatives to which C_N is applied.) Call this the case of *exogenous* (motivating or normative) reasons. We obtain a more sophisticated, and perhaps more realistic, picture of an agent's choice dispositions by allowing the set of (motivating or normative) reasons to depend on which alternatives are available. In other words, different sets of available alternatives may *endogenously* activate different sets of (motivating or normative) reasons.

Recall our example of an agent choosing between three different types of cars. If the agent's set of motivating reasons, which could be any subset of 'the car is fast' (F), 'it is big' (B), 'it is environmentally friendly' (E), were exogenously given,

everything would be as in our example of choices between fruits: the choice function over cars would look much like the one over fruits, just induced by the appropriate preference order over cars instead of the one over fruits. But if different sets of available alternatives somehow activated different sets of motivating reasons, the agent's choice dispositions would be more complex. Suppose, for instance, that a choice between any two cars leads the agent to be motivated by all and only those reasons that distinguish those cars. The agent may then exhibit 'cyclical' choice dispositions. In our example, he or she would choose the Sports Beetle ($f\neg be$) over the Family Hybrid ($\neg fbe$) when presented with two environmentally friendly cars, the Family Hybrid ($\neg fbe$) over the Monster Hummer ($fb\neg e$) when presented with two big cars, and the Monster Hummer ($fb\neg e$) over the Sports Beetle ($f\neg be$) when presented with two fast cars. The resulting choice function would not be explicable in terms of any single preference order, since any such order would have to rank the Sports Beetle over the Family Hybrid, the Family Hybrid over the Monster Hummer, and yet the Monster Hummer over the Sports Beetle, a violation of transitivity.

While standard rational choice theory is unable to explain those kinds of choice dispositions, let alone to rationalize them, our theory can account for them by pointing to the way the reasons for those choices are affected by the available alternatives – or more generally, by the choice context. Such context dependence of preferences and choices is widely recognized and seen by many as a serious challenge to rational choice theory. By making rational choice theory reason-based as suggested in this paper, we have therefore introduced some new conceptual resources for analyzing this phenomenon as well.

13 Concluding remarks

We have proposed a reason-based theory of rational choice, which responds to the widely held concern that standard rational choice theory does not say anything about the reasons underlying preferences but holds preferences to be unchangeable and not subject to reason-based scrutiny. Our theory can be viewed from two angles. On the one hand, it generalizes standard rational choice theory and thereby connects with the large body of work in decision theory and the social sciences on the formal modelling of human decision problems. On the other hand, it formalizes the role of reasons in rational decision making, thereby capturing a core concern of the philosophical literature on the relationship between reasons and actions.

Our theory should not be regarded as a rival to either body of work. Instead, our aim is to promote a dialogue between formal rational choice theory and philosophical work on reasons. Although we have only presented a first sketch of our theory and much further work is needed, we hope that the concepts and tools provided in this paper will help to advance this enterprise.

A   Appendix

We here prove Theorems 1 and 2. The proofs of Theorems 1* and 2* – the extensions to preferences over uncertain prospects – are significantly more technical and are available on request. Throughout our proofs, we write $M_x := \{R \in M : R$ is true of $x\}$ to denote the combination of reasons in a given set $M \subseteq \mathcal{P}$ that are true of an alternative x in X, and we write $\mathcal{S} := \{S \subseteq \mathcal{R} : S$ is consistent$\}$ to denote the set of all possible combinations of reasons.

The proof of both theorems uses the following lemma, which holds independently of any assumptions on the set $\mathcal{M}$ of possible motivating sets.

Lemma 1. Suppose Axiom 1 holds. For all x, y, x', y' in X and all M in $\mathcal{M}$, if $\{R \in M : R$ is true of $x\} = \{R \in M : R$ is true of $x'\}$ and $\{R \in M : R$ is true of $y\}$ $= \{R \in M : R$ is true of $y'\}$ then $x \succsim_M y \Leftrightarrow x' \succsim_M y'$.

Proof. Let $x, y, x', y' \in X$ and $M \in \mathcal{M}$ such that $M_x = M_{x'}$ and $M_y = M_{y'}$. Applying Axiom 1 twice, we have $x \sim_M x'$ and $y \sim_M y'$. So, as $\succsim_M$ is transitive, $x \succsim_M y \Leftrightarrow x' \succsim_M y'$. ∎

A.1   Proof of theorem 1

We first prove necessity and then sufficiency of our two axioms for the representation of preferences in terms of a weighing relation.

A.1.1   Necessity of the axioms for the representation

First, suppose a binary relation $\geq$ on $\mathcal{S}$ generates all preference orders $\succsim_M$ across $M \in \mathcal{M}$. Axiom 2 is obviously satisfied. As for Axiom 1, consider any $M \in \mathcal{M}$ and any $x, y \in X$ such that $M_x = M_y$. We have to show that $x \sim_M y$. As $\succsim_M$ is reflexive, we have $x \sim_M x$. So, since $\geq$ generates $\succsim_M$, we must have $M_x \equiv M_x$. But since $M_x = M_y$, this implies $M_x \equiv M_y$. From this – again using the fact that $\geq$ generates $\succsim_M$ – it follows that $x \sim_M y$, as required.

A.1.2   Sufficiency of the axioms for the representation

Now assume that Axioms 1 and 2 are satisfied. Recall that $\mathcal{M}$ is closed under finite intersection. This is part (i) of our regularity assumption on $\mathcal{M}$. There is no need to assume part (ii) for Theorem 1.

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

To prove this claim, let $x, y, x', y' \in X$ and $M, M' \in \mathcal{M}$ with $M_x = M'_{x'}$ and $M_y = M'_{y'}$. As $\mathcal{M}$ is closed under finite intersection, we have $M \cap M' \in \mathcal{M}$. We first

show that
$$(M \cap M')_x = (M \cap M')_{x'} = M_x = M'_{x'},$$
$$(M \cap M')_y = (M \cap M')_{y'} = M_y = M'_{y'}.$$

To see that the first set of identities holds, notice the following: firstly, $M_x = M'_{x'}$ by assumption; secondly, $(M \cap M')_x = M_x$, since $(M \cap M')_x = M_x \cap M'_x = M_x$ (the last identity holds because $M'_x \supseteq (M'_{x'})_x = (M_x)_x = M_x$); and, thirdly, $(M \cap M')_{x'} = M'_{x'}$, since $(M \cap M')_{x'} = M_{x'} \cap M'_{x'} = M'_{x'}$ (the last identity holds because $M_{x'} \supseteq (M_x)_{x'} = (M'_{x'})_{x'} = M'_{x'}$). The second set of identities holds by an analogous argument.

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

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

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

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

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

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

The equivalences (*) to (***) together imply that $x \succsim_M y \Leftrightarrow x' \succsim_{M'} y'$. ∎

Claim 1 allows us to define a binary relation $\geq$ on $\mathcal{S}$ with the following properties: 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 each $M \in M$, $\geq$ generates $\succsim_M$, that is, $x \succsim_M y \Leftrightarrow M_x \geq M_y$ for all $x, y \in X$.

To prove this claim, let $M \in \mathcal{M}$ and $x, y \in X$. First, assume $x \succsim_M y$. We show that $M_x \geq M_y$, that is, $x' \succsim_{M'} y'$ for some $x', y' \in X$ and $M' \in \mathcal{M}$ with $M'_{x'} = M_x$ and $M'_{y'} = M_y$. This obviously holds: simply take $x' = x$, $y' = y$, and $M' = M$. Conversely, assume that $M_x \geq M_y$. Then, by the definition of $\geq$ and Claim 1, we have $x' \succsim_{M'} y'$ for *all* $x', y' \in X$ and $M' \in \mathcal{M}$ satisfying $M'_{x'} = M_x$ and $M'_{y'} = M_y$. In particular, $x \succsim_M y$. This completes the proof of Theorem 1. ∎

A.2 Proof of theorem 2

Assume a weakly independent set of possible reasons $\mathcal{P}$. The proof is written so as to maximize parallels with the proof of Theorem 1.

A.2.1 Necessity of the axioms for the representation

By the argument in the earlier proof, Axioms 1 and 2 hold if some order $\geq$ on $\mathcal{S}$ generates all preference orders $\succsim_M$, across $M \in \mathcal{M}$.

A.2.2 Sufficiency of the axioms for the representation

Now assume that Axioms 1 and 2 are satisfied. Recall that, for any M, M' in $\mathcal{M}$, $\mathcal{M}$ contains some superset of $M \cup M'$. This is a weakened variant of part (ii) of our regularity assumption on $\mathcal{M}$. There is no need to assume part (i) for Theorem 2.

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

This claim, although analogous to the first claim in the proof of Theorem 1, requires a different proof. Let $x, y, x', y' \in X$ and $M, M' \in \mathcal{M}$ such that $M_x = M'_{x'}$ and $M_y = M'_{y'}$. As $\mathcal{M}$ contains M, M', it contains some $M'' \supseteq M \cup M'$, by assumption. By the weak independence of $\mathcal{P}$, there are $a, b \in X$ such that $\mathcal{P}_a = M_x$ and $\mathcal{P}_b = M_y$. Hence $M''_a = M_x$ and $M''_b = M_y$, so that by Axiom 2

$$a \succsim_{M''} b \Leftrightarrow a \succsim_M b. \tag{*}$$

By an analogous argument (performed on x', y', M' instead of x, y, M), there are $a', b' \in X$ such that $M''_{a'} = M'_{x'}$ and $M''_{b'} = M'_{y'}$ and

$$a' \succsim_{M''} b' \Leftrightarrow a' \succsim_{M'} b'. \tag{**}$$

Using Lemma 1, the right-hand side of (*) is equivalent to $x \succsim_M y$ (because $M_a = M_x$ and $M_b = M_y$); the right-hand side of (**) is equivalent to $x' \succsim_{M'} y'$ (because $M'_{a'} = M'_{x'}$ and $M'_{b'} = M'_{y'}$); and the left-hand sides of (*) and (**) are equivalent to each other (because $M''_a = M''_{a'}$ and $M''_b = M''_{b'}$). These three equivalences together with the equivalences (*) and (**) imply that $x \succsim_M y \Leftrightarrow x' \succsim_{M'} y'$. ∎

Claim 1 allows us to define a binary relation $\geq^*$ on $\mathcal{S}$, analogous to the one defined in our proof of Theorem 1. But this time it is only a precursor to the relation we ultimately wish to define (it must subsequently be extended to an order). The relation $\geq^*$ has the following properties: for any $S, S' \in \mathcal{S}$, $S \geq^* S'$ 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 each $M \in M$, the binary relation $\geq^*$ generates $\succsim_M$, that is, $x \succsim_M y \Leftrightarrow M_x \geq M_y$ for all $x, y \in X$.

The proof is analogous to that of the second claim in the proof of Theorem 1. ∎

Claim 3. $\geq^*$ is transitive.

Consider $S, S', S'' \in \mathcal{S}$ such that $S \geq^* S'$ and $S' \geq^* S''$. We have to show that $S \geq^* S''$. Since $S \geq^* S'$, there exist $x, y \in X$ and $M \in \mathcal{M}$ such that $M_x = S$, $M_y = S'$ and $x \succsim_M y$. Since $S' \geq^* S''$, there exist $y', z \in X$ and $M' \in \mathcal{M}$ such that

$M'_{y'} = S'$, $M'_z = S''$ and $y' \succsim_{M'} z$. Since $M, M' \in \mathcal{M}$ and by our assumption on $\mathcal{M}$, $\mathcal{M}$ contains some $M'' \supseteq M \cup M'$. By the weak independence of $\mathcal{P}$, there are $a, b, c \in X$ such that $\mathcal{P}_a = S$, $\mathcal{P}_b = S'$ and $\mathcal{P}_c = S''$, whence $M''_a = S$, $M''_b = S'$ and $M''_c = S''$. Since $x \succsim_M y$, $M_x = M''_a \ (= S)$, and $M_y = M''_b \ (= S')$, and by Claim 1, we have $a \succsim_{M''} b$. Similarly, since $y' \succsim_{M'} z$, $M'_{y'} = M''_b \ (= S')$, and $M'_z = M''_c$ $(= S'')$, and by Claim 1, we have $b \succsim_{M''} c$. Since $a \succsim_{M''} b$ and $b \succsim_{M''} c$, and by the transitivity of $\succsim_{M''}$, we have $a \succsim_{M''} c$. So, by the definition of $\geq^*$ (and using the fact that $M''_a = S$ and $M''_c = S''$), we have $S \geq^* S''$. ∎

Claim 4. There exists an order $\geq$ on S that extends $\geq^*$, in the usual sense that $S >^* S' \Rightarrow S > S'$ and $S \equiv^* S' \Rightarrow S \equiv S'$ for all $S, S' \in S$; equivalently, $S \geq S' \Leftrightarrow S \geq^* S'$ for all $S, S' \in S$ that are ranked relative to each other by $\geq^*$.

This follows from Claim 3 via a classic extension theorem for binary relations.[39]
∎

Claim 5. For each $M \in M$, the order $\geq$ defined in Claim 4 generates $\succsim_M$.

To prove this claim, let $M \in \mathcal{M}$ and $x, y \in X$. First, if $x \succsim_M y$, then $M_x \succsim^* M_y$, as $\geq^*$ generates $\succsim_M$ (by Claim 2), whence $M_x \geq M_y$, as $\geq$ extends $\geq^*$. Conversely, if $x \not\succsim_M y$, then $y \succ_M x$ (as $\succsim_M$ is complete), so that $M_y >^* M_x$, as $\geq^*$ generates $\succsim_M$ (by Claim 2). This implies that $M_y > M_x$, since $\geq$ extends $\geq^*$, hence that $M_x \not\geq M_y$. This completes the proof of Theorem 2. ∎

[39] As proven in its most general form by K. Suzumura, 'Remarks on the theory of collective choice', *Economica* 43 (1976), pp. 381–90.