Representation Theorems and the Grounds of Intentionality

EDWARD ELLIOTT

A thesis submitted for the degree of
Doctor of Philosophy of
The Australian National University
August 2015

Statement

This thesis is solely the work of its author. No part of it has previously been submitted for any degree, or is currently being submitted for any other degree. To the best of my knowledge, any help received in preparing this thesis, and all sources used, have been duly acknowledged.

EDWARD ELLIOTT
12th August, 2015

Acknowledgements

This thesis has benefitted greatly from the input of others. I particularly want to thank the Chair of my supervisory panel, David Chalmers, who has helped me to weed out bad ideas and, moreover, to turn what's left into something worthwhile. Thanks are also due to other members of my panel: Wolfgang Schwarz (who was an early source of many of my ideas on representation theorems), Hanti Lin, and especially Alan Hájek for sharp comments on my writing, and Rachael Briggs for enthusiastically reading and critiquing numerous early drafts.

I am also grateful to those who have helped to make my years at ANU enjoyable and philosophically enlightening. My work on this thesis has benefitted directly from talking with Rachael Brown, Dave Kalkman, Ole Koksvik, Dan Korman, Angela Mendelovici, Daniel Nolan, Alexander Sandgren, Paolo Santorio, Jon Simon, and Kim Sterelny. A special thanks to Leon Leontyev for many discussions (and debates) on formal epistemology and decision theory; to Kelvin McQueen and Clas Weber for countless hours of discussion on two-dimensional semantics and the foundations of meaning; and to Jessica Isserow for lively discussions on every philosophical question, ever. Thanks also to audiences here at the ANU, at annual meetings of the Australasian Association of Philosophy, and numerous Kioloas.

On a more personal note, I would like to thank my parents for their love, constant encouragement, and support, without which these years would have been much more difficult. I would also like to thank Wilumpa Jarern-Iamsakul for her support during the first few years. And finally, I would like to thank Jessica Isserow for her friendship, love, and support—since I met you, you've been the best friend I've ever had, and I couldn't have finished this without you.

Abstract

This work evaluates and defends the idea that decision-theoretic representation theorems can play an important role in showing how credences and utilities can be characterised, at least in large part, in terms of their connection with preferences. Roughly, a *decision-theoretic representation theorem* tells us that if an agent's preferences satisfy constraints C, then that agent can be represented as maximising her expected utility under a unique set of credences (modelled by a credence function $\mathcal{B}el$) and utilities (modelled by a utility function $\mathcal{D}es$). Such theorems have been thought by many to not only show how credences and utilities can be understood *via* their relation to preferences, but also to show how credences and utilities can be *naturalised*—that is, characterised in wholly non-mental, non-intentional, and non-normative terms.

There are two broad questions that are addressed. The first (and more specific) question is whether any version of characterisational representationism, based on one of the representation theorems that are currently available to us, will be of much use in directly advancing the long-standing project of showing how representational mental states can exist within the natural world. I answer this first question in the negative: no current representation theorem lends itself to a plausible and naturalistic interpretation suitable for the goal of reducing facts about credences and utilities to a naturalistic base. A naturalistic variety of characterisational representationism will have to await a new kind of representation theorem, quite distinct from any which have yet been developed.

The second question is whether characterisational representationism in any form (naturalistic or otherwise) is a viable position—whether, in particular, there is any value to developing representation theorems with the goal of characterising what it is to have credences and utilities in mind. This I answer in the affirmative. In particular, I defend a weak version of characterisational representationism against a number of philosophical critiques. With that in mind, I also argue that there are serious drawbacks with the particular theorems that decision theorists have developed thus far; particularly those which have been developed within the four basic formal frameworks developed by Savage, Anscombe and Aumann, Jeffrey, and Ramsey.

In the final part of the work, however, I develop a new representation theorem, which I argue goes some of the way towards resolving the most troubling issues associated with earlier theorems. I first show how to construct a theorem which is ontologically similar to Jeffrey's, but formally more similar to Ramsey's—but which does not suffer from the infamous problems associated with Ramsey's notion of ethical neutrality, and which has stronger uniqueness results than Jeffrey's theorem. Furthermore, it is argued that the new theorem's preference conditions are descriptively reasonable, even for ordinary agents, and that the credence and utility functions associated with this theorem are capable of representing a wide range of non-ideal agents—including those who: (i) might have credences and utilities only towards non-specific propositions, (ii) are probabilistically incoherent, (iii) are deductively fallible, and (iv) have distinct credences and utilities towards logically equivalent propositions.

A Note on Notation

Throughout this thesis, I have maintained a consistent notational scheme, which I have summarised here for convenience. Sections where the relevant notions are introduced and discussed are included in the parentheses.

f	Arbitrary function
$\mathcal{B}el$	Function intended to represent credences (§2.1, §2.5)
$\mathcal{D}es$	Function intended to represent utilities (§2.1, §2.5)
$\mathcal{P}r$	Probability function (need not be intended to represent credences) (Definition 2.2)
$\mathcal{EU}$	Expected utility function (§2.4)
$\succcurlyeq$	Preference relation (Definition 2.5)
$\succcurlyeq^{b}$	Relative credence relation (Definition 2.8)
$\succcurlyeq^{x}$	Arbitrary binary relation
$\mathcal{X} = \{x, y, z, \ldots\}$	Arbitrary set
$\mathcal{W} = \{w_1, w_2, w_3, \ldots\}$	Set of *possibilities*; usually a set of *possible worlds*
$\mathcal{P} = \{P, Q, R, \ldots\}$	Set of *propositions*; in some cases a set of subsets of $\mathcal{W}$
$\mathcal{S} = \{s_1, s_2, s_3, \ldots\}$	Set of *states*; a partition of some possibility space $\mathcal{W}$ (§5.1.1)
$\mathcal{E} = \{E_1, E_2, E_3, \ldots\}$	Set of *events*; i.e., a set of subsets of $\mathcal{S}$ (§5.1.1)
$\mathcal{O} = \{o_1, o_2, o_3, \ldots\}$	Set of *outcomes*; usually a set of propositions (§5.1.1)
$\mathcal{A}’ = \{\alpha, \beta, \gamma, \ldots\}$	Set of *acts* (§5.1.1), or *intentions to act* (§5.4)
$\mathcal{A} = \{\mathcal{F}, \mathcal{G}, \mathcal{H}, \ldots\}$	Set of *act-functions*; i.e., functions from some $\mathcal{S}' \subseteq \mathcal{S}$ into some $\mathcal{O}$ (§5.1.1, Definition 5.4, Appendix B)
$\mathcal{M}_{\mathcal{X}} = \{\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3, \ldots\}$	Set of all *lottery-functions* on a set $\mathcal{X}$; i.e., a set of functions from $\mathcal{X}$ into [0, 1] (Definition 6.1, Definition 6.2)
$\mathcal{H} = \{\hbar_1, \hbar_2, \hbar_3, \ldots\}$	Set of *horse-functions*; i.e., functions from some $\mathcal{S}' \subseteq \mathcal{S}$ into some $\mathcal{M}_{\mathcal{O}}$ (§6.1.1)
$\mathcal{G} \subseteq \mathcal{O} \times \mathcal{P} \times \mathcal{O}$	Set of (two-outcome) *gambles*; members usually represented $(o_1, P; o_2)$ (§7.1, §8.1.1)
$\mathcal{N}$	Set of *null events* (Definition 5.10) or *null propositions* (Definition 6.4)

Contents

Chapter one

Beliefs, Credences, and the Naturalisation of Intentionality

In his ‘Radical Interpretation’ (1974), David Lewis sets us a challenge: an ordinary person named Karl is the subject of our investigation, and the task is to determine what he believes and desires without presupposing any particular claims to that effect. We have at our disposal all the facts we could ever want about Karl—about his upbringing, neurobiological constitution, ancestral history, and external societal context—*except* for those facts which directly inform us as to the contents of his beliefs and desires. The challenge seems in principle satisfiable; it’s unlikely that it’s a brute fact about Karl that he believes and desires as he does, so if he *believes that P* or *desires that Q*, such things should supervene on other truths which are not directly about his attitudes.

Beliefs and desires are centrally important to the folk conception of the mind, so it would be very useful to develop a non-circular characterisation of when a subject *believes that P* and *desires that Q*. We can refer to this as a *characterisation project*; it is, by all accounts, still very much incomplete. But Lewis—like many others who have accepted the same challenge—engages upon a yet more ambitious project still: to *naturalise* beliefs and desires, by accounting for what it is to be in such states whilst appealing only to *non-intentional, non-mental,* and *non-normative* factors. We can refer to this a *naturalisation project*—it is an instance of a characterisation project, with an added twist. As a physicalist, Lewis took his task to be the explanation of beliefs and desires in entirely physical terms: “Given **P**, the facts about Karl as a physical system, solve for the rest” (1974, 331).[1] For those engaged in the naturalisation project, it is not enough to just say in non-cognate terms what it is to have beliefs and desires; rather, we need to show how these attitudes fit within the normal causal order of physical objects and natural properties. Beliefs and desires are intentional states—they are *about* things—and intentionality just does not seem to be a metaphysically fundamental phenomena. Jerry Fodor nicely sums up the intuitions here:

[1] Lewis also hoped to supply a naturalistic account of what Karl means by the terms and sentences he uses, discussion of which would take us well beyond the scope of this work.

> I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm*, and *charge* will perhaps appear on their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. (1987, 97)

We ought, it seems, to be able to explain how the intentionality arises within a natural, and fundamentally non-intentional, world.

I have so far cast these characterisation and naturalisation projects in terms of beliefs and desires, but like the rest of us, Karl presumably has *partial beliefs* and different *strengths of desire* as well. As a well-informed individual, Karl is likely more confident that evolutionary theory is broadly correct than he is in the predictive accuracy of astrology. When deciding upon an eatery, he feels a stronger desire for Turkish cuisine than for Korean. Furthermore, we appeal to these differences of degree when explaining Karl's behaviour—e.g., in explaining why he opted for Turkish, and why he ignores the astrology section of the newspaper. Indeed, inasmuch as we are already willing to accept that Karl has propositional attitudes *at all*, it seems that we can take it for granted that some of these attitudes come in degrees. At the very least, ordinary agents have *credences* (i.e., degrees of belief) and *utilities* (i.e., strengths of desire). Gradation is an important—perhaps ineliminable—part of the ordinary, folk conception of the mind, and if we are not eliminativists about the folk categories of belief and desire then we ought not to be eliminativists about credences and utilities either.

Historically, philosophers who have engaged in anything like a characterisation or naturalisation project have centred their attention on beliefs and, to a lesser extent, the other non-graded propositional attitudes.[2] Progress on this front, however, has been slow in recent decades. The end of the 20th Century saw the development of the *representational theory of mind* (Putnam 1980, and esp. Fodor 1975), which (roughly) takes propositional attitudes to be relations that a subject bears to representational structures stored somewhere in the head that play a particular kind of computational role. With that theory came a large amount of work on the naturalisation of conceptual content. However, the three most influential views which have been developed—namely, the *causal-informational*

[2] Like any other 'big projects' in philosophy, there will be some who see naturalisation projects and the broader characterisation projects as misguided from the outset. Though it is very much a minority position, some believe that intentional and semantic facts may be metaphysically basic (Kearns and Magidor 2012). Others might deny that propositional attitudes of the kind that folk psychology refers to even exist (Quine 1960, Churchland 1981, Stitch 1992), or that credences may be (finitely) indefinable (as Eriksson and Hájek 2007 seem to advocate). A response to these positions is beyond the scope of this work—one has to start somewhere, and a realist, non-primitivist take on the graded attitudes seems about as good as any. I am sympathetic to the idea that credences and utilities cannot be given a fully naturalistic *and finite* characterisation, but the best argument for that position rests on the consistent failure of attempts to provide such a characterisation, and we are hardly at the stage where giving up is warranted.

theories of Dretske (1981) and Fodor (1987), *teleosemantics* (Millikan 1989), and *conceptual role semantics* (Block 1986)—still suffer from unresolved problems recognised since their inception. Worse still, it is unclear whether and how these approaches to understanding belief—including the background theory of the propositional attitudes—might be augmented or modified to accommodate the graded attitudes.[3] Major alternative approaches (e.g., Davidson 1973, Stalnaker 1984, Dennett 1989) retain the emphasis on belief—allusions are sometimes made towards an account of credences and utilities, but details are sparse or non-existent.

A shift in focus may prove helpful: perhaps progress is to be found by accounting *first* for the graded attitudes, and then seeing what this might teach us about the nature of the non-graded attitudes—if beliefs and desires are still considered important once we have accounted for credences and utilities (cf. Jeffrey 1970, Christensen 2004). Indeed the focus on beliefs in particular is surprising given the long-standing and common view that we should ultimately understand beliefs in terms of credences.[4] If something like this is true, then our attention should presumably be directed in the first instance towards an understanding of the graded attitudes. And even if it is not true, both the characterisation and naturalisation projects as applied to credences and utilities (as opposed to beliefs and desires) are independently important ventures worthy of philosophical attention.

In general, philosophers have had much to say on what our credences and utilities *should* be like at a given time, how we should change them over time in response to learning, and how they ought to influence our decisions. We read, for example, that our credences ought to obey certain ‘coherence’ conditions, and that our utilities shouldn’t be intransitive. And as just noted, there is also a limited amount of work on how credences relate to beliefs. But very little has been said on just what credences and utilities *are*, on *what it is* for an agent to have the credences and utilities that they do.

More specifically, the issue here is to characterise the conditions under which an agent counts as being in such-and-such a credence or utility state. It is orthodox to hold that beliefs and desires are binary relations between a subject at a time and a proposition. I suspect that most would be happy to say something similar about the graded attitudes. That is, it seems plausible to say that a *credence of x in P* is a ternary relation between a subject at a time, a degree (represented by some value, x), and the proposition P (cf. Huber 2009, 2). Likewise for utilities. The real philosophical meat lies in specifying the conditions under which an agent stands in such a relationship—and whether those conditions can be stated in entirely naturalistic terms.

[3] I discuss these issues further in Chapter 4.

[4] See (Eriksson and Hájek 2007, 206-7) for reasons in favour of this view. One idea here is the straightforward version of the Lockean thesis, that to believe that P is simply to have a sufficiently high credence in P (see Foley 1992, Hawthorne 2009). But there are many other options for reducing beliefs to credences (e.g., Weatherson 2005, 2012a, forthcoming, Clarke 2013).

A small amount of discussion does exist on attempts to characterise credences and utilities. As a rule, inasmuch as philosophers have engaged with this project at all, the general strategy has been to explain these attitudes more or less wholly in terms of *preferences*. The famous *betting interpretation*, despite numerous critiques, remains a perennially popular instance of this strategy (for recent defences, see Williamson 2010, Shafer 2011). Likewise for some *interpretivist* views, according to which, as Donald Davidson puts it, "Subjective probabilities [i.e., credences] and quantified desires [i.e., utilities] are … theoretical constructs whose function is to relate and explain simple preferences" (2004).[5]

Exactly what is meant by 'preference' here is something we will discuss in more detail below (see §2.2). A common idea, reasonably standard amongst those whose work centres around decision and game theory, is that preferences can be understood *behaviourally*: an agent *S*'s preference ranking is supposed to more or less directly encode her behavioural dispositions (Samuelson 1938), or at least possess very strong links to those dispositions (as suggested for example by Davidson 1990, 317), such that we can determine *S*'s preferences given sufficient observation of her behaviour. If this is true, then a characterisation of credences and utilities in terms of preferences points the way to a naturalisation of those attitudes. This goal, of course, is often in the background (and sometimes in the foreground) as a key motivation behind preference-based approaches to characterising credences and utilities. To the extent that the notion of preference is itself naturalistically kosher, or at least more directly amenable to naturalisation than the intentional states with which we began, then, *if* we could characterise credences and utilities in terms of preferences, we would be well placed to supply a fully naturalistic account of two important—and perhaps even *basic*—intentional attitudes.[6]

To many, the most promising path for developing a preference-based characterisation of credences and utilities involves the appeal to one or another of the numerous *representation theorems* which have been developed for *classical expected utility* (CEU) theory. To gloss over several important details, which we will return to below, these theorems are generally taken to imply something along the following lines:

> If an agent's preferences satisfy constraints *C*, then that agent can be *represented* as maximising her expected utility under a unique set of credences (modelled by a probability function $\mathcal{B}el$) and utilities (modelled by a utility function $\mathcal{D}es$)

[5] See §4.2 for more discussion on interpretivism, including versions of the position which do not characterise credences and utilities *solely* in terms of preferences.

[6] To be clear, there are many kinds of intentional states, so to account for just credences and utilities is by no means to account for intentionality across the board. It may turn out, however, to be a particularly important part of the overall naturalisation project—especially if beliefs and desires reduce to graded attitudes, and linguistic meanings depend on speakers' attitudes.

If we assume that ordinary agents satisfy C, and somehow establish a close connection between how the agent can be *represented* and how she in fact *is*, then we appear to be well on our way to giving a preference-based account of what credences and utilities *are*: $\mathcal{B}el$ represents her credences, $\mathcal{D}es$ represents her utilities, and to have credences $\mathcal{B}el$ and utilities $\mathcal{D}es$ is just to have preferences which can be represented as such. At least, that is a very straightforward (and ultimately very flawed) version of the idea—as we will see, there are plenty of modifications to be made, but the gist of the view should be clear enough for now. In this sense, representation theorems are often seen as playing a central role in the *conceptual foundations* of decision theory and epistemology: they help to characterise the very notions that decision theorists and epistemologists are theorising about.

Representation theorems also exist for so-called *non-classical utility* (NCU) theories of decision-making, or theories which deviate from the classical expected utility norms.[7] Examples of such theories include, amongst many others, *cumulative prospect theory* (Tversky and Kahneman 1992), *weighted utility theories* (Fishburn 1983), *Choquet expected utility models* (Schmeidler 1989), *risk-weighted expected utility theory* (Buchak 2013), and *maxmin expected utility theory* (Alon and Schmeidler 2014). The potential application of NCU theorems in the preference-based characterisation of credences and utilities has been largely ignored by philosophers to date, though it is a lively project within other disciplines.

As will become clear below, I doubt that that credences and utilities can be understood *solely* in terms of preferences. However, let us use *preference functionalism* for the weaker (and more plausible) view that credences and utilities ought to be characterised *at least in large part* in terms of preferences. This view has a lot more going for it: the standard for the past few decades has been to understand mental states in terms of what they *do*—and if credences and utilities are supposed to do *anything*, they are involved in the explanation of our preferences. It is worth noting also that preference functionalism is not an inherently anti-realist or behaviourist position (cf. §4.5); it should be treated as neutral with respect to whether credences *are* preference states, or if credences are to be functionally characterised (partly) in terms of the preference patterns they *give rise to*.

Furthermore, let us use *characterisational representationism* for the particular variety of preference functionalism whereby decision-theoretic representation theorems are taken to play a centrally important role in showing how credences and utilities can ultimately be characterised, at least in large part, in terms of preferences. (Characterisational representationism will be discussed in depth in Chapter 3 and Chapter 4.) For philosophical discussions friendly to the view, see (Ramsey 1931), (Savage 1954), (de Finetti 1964, 1974), (Anscombe and Aumann 1963), (Harsanyi 1977), (Eells 1982), (Jeffrey 1968, 1990), (Davidson 1980, 1990, 2004), (Pettit 1991), (Maher 1993, 1997), and (Schwarz 2014b). Representation theorems also appear to be in the background of Lewis' own

[7] NCU theorems and CEU theorems are distinguished more thoroughly in §2.4.

sketch for deriving the intentional facts about Karl from the set of all basic physical facts (see §4.2).

Characterisational representationism is the most common variety of preference functionalism amongst contemporary philosophers, and a common view *simpliciter*—so much so that Colin Howson and Peter Urbach describe it as having become "so dominant … that it is fair to call it now the *orthodox account*" (2005, 57). This is especially true in economics and decision-theoretic psychology, where something like characterisational representationism is more or less an unquestioned orthodoxy. The influence of the position holds even despite a number of recent sceptical discussions; e.g., (Hampton 1994), (Joyce 1999, Ch. 3), (Christensen 2001, 2004), (Howson and Urbach 2005, Ch. 3), (Eriksson and Hájek 2007), (Easwaran 2014), (Dogramaci forthcoming), and especially (Meacham and Weisberg 2011). The main worries raised by these authors are discussed below, where I argue that they don't give us sufficient grounds for rejecting characterisational representationism *tout court*—though they do give us reasons to reject very strong and simplistic versions of the view.

In this work, I will evaluate the status of characterisational representationism. There are two main questions that I want to address. The first (and more specific) question is whether characterisational representationism will be of much use in directly advancing the naturalisation project, given the theorems that we currently have available—that is, whether we might appeal to any of the representation theorems we have now in providing an entirely non-intentional and non-mental account of what it is to have such-and-such credences and utilities.

I answer this first question in the negative: no current representation theorem lends itself to a plausible and naturalistic interpretation suitable for the goal of reducing facts about credences and utilities to a naturalistic base. My argument for this, moreover, is not grounded in concerns over the philosophical merits of (pseudo-)behaviourism or anti-realist construals of propositional attitudes, which have motivated much of the scepticism that has been directed towards characterisational representationism. Most representation theorems simply don't lend themselves well to a naturalistic interpretation, and where they do, it is a mistake to think that they can be given a behavioural or otherwise non-intentional interpretation *inasmuch as* their $\mathcal{B}el$ and $\mathcal{D}es$ functions are to plausibly model decision-makers' credences and utilities. As a consequence of how *objects of choice* are formalised in our current systems, a naturalistic interpretation of any current theorem—to whatever extent it may exist—comes at the cost of breaking any plausible connection between the established representation and the mental facts of the matter. Furthermore, the most general framework we have for connecting credences and utilities to behaviour is incapable of capturing those attitudes for a very wide range of important propositions (towards which we almost certainly do have credences and utilities). A naturalistic variety of characterisational representationism will have to await a new kind of representation theorem, quite distinct from any which have yet been developed.

The second question is whether characterisational representationism in any form (naturalistic or otherwise) is a viable position—whether, in particular, there is any value to developing representation theorems with the goal of characterising what it is to have credences and utilities in mind. This I answer in the affirmative. In particular, I defend a weak version of characterisational representationism against a number of philosophical critiques. With that in mind, I also argue that there are serious drawbacks with the particular theorems that decision theorists have developed thus far. In the final part of the work, however, I develop a new representation theorem, which I argue goes *some* of the way towards resolving the most troubling issues associated with earlier theorems.

1.1 Structure of the discussion

In the next chapter, I will introduce and clarify the technical concepts and vocabulary used throughout the rest of the thesis, including: a number of formal models for the representation of credences, representation theorems, uniqueness theorems, and the interpretations thereof.

In Chapters 3 and 4, I give a partial defence of characterisational representationism from a number of sceptical critiques found in the recent philosophical literature. Chapter 3 discusses in detail a common but naïve version of characterisational representationism—what I call the *classical* theory—and then looks at where it goes wrong. The biggest concern with the classical theory is that it takes an anti-realist stance towards credences and utilities, treating them as mere redescriptions of preference patterns rather than independently existing mental states in their own right. Other major concerns stem from the particular kinds of representation theorems that have traditionally been appealed to—*viz.*, CEU theorems, developed for primarily normative purposes. The final section of Chapter 3 outlines a number of desiderata that a representation theorem ought to satisfy if it is underlie a plausible version of characterisational representationism.

Then, in Chapter 4, I argue that with *the right kind of representation theorem*—one which satisfies the stated desiderata—the central worries with classical characterisational representationism might be overcome. Indeed, when placed in comparison with alternatives, a more sophisticated version of characterisational representationism based on an appropriate theorem has distinct advantages which should make it attractive to philosophers seeking to understand the nature of our graded propositional attitudes. In particular, I argue that given the right theorem, characterisational representat-ionism should seem especially promising in helping us to pin down the intentional content of these attitudes.

This is followed by a review of a large number of representation theorems in decision theory, with a focus on their viability as foundations for characterisational representationism (naturalistic or otherwise). This is done in light of the desiderata developed in Chapter 3. It is, in other words, an enquiry into whether *the right kind of representation theorem* currently exists.

Chapter 5 focuses on Savage's theorem and the formal paradigm that he created; there, I find that the reliance on functions from states to outcomes (constituting these theorems' formal representations of *acts*) leads to deep problems which limit the usefulness of all Savagean theorems—both for characterisational representationism and more generally. I also consider and reject the feasibility of a purely naturalistic understanding of acts and preferences, two basic notions involved in the interpretation of Savage-like theorems. Chapter 6 then considers two other broad classes of representation theorem: the lottery-based framework (found in the theorems of von Neumann and Morgenstern, and Anscombe and Aumann) and the monoset framework (found in the Bolker-Jeffrey theorem). These theorems, too, are found wanting, though for very different reasons. Finally, Chapter 7 evaluates Ramsey's representation theorem. The well-known problem of ethical neutrality is raised, and it is argued that Ramsey's assumption of the existence of ethically neutral propositions is not a mere idealisation that can be simply overlooked.

Between them, Chapters 5 through to 7 cover the vast majority of representation theorems that have been developed over the past century. Jointly, they demonstrate that these theorems are not up to the task of founding a plausible and complete version of characterisational representationism. There are five broad kinds of problems that arise, centred on the following themes:

1. *Satisfiability*: whether a theorem T's preference conditions (under a reasonable interpretation) are satisfied (or approximately satisfied) by ordinary agents.
2. *Plausibility*: whether, supposing that S satisfies T's preference conditions, the resulting model of S's credences, utilities, and decision-making procedure is intuitively and empirically plausible.
3. *Uniqueness*: whether the model of S's credences is, in an interesting sense, unique.
4. *Circularity*: whether any useful decision-theoretic interpretation of T depends on a prior specification of S's credences and utilities.
5. *Naturalisability*: whether the decision-theoretic interpretation of T involves an unavoidable appeal to some intentional state or other.

Issues surrounding the 'naturalisability' of a theorem are, of course, only applicable to those engaged in the naturalisation project. Problems regarding 'circularity' of course imply problems of 'naturalisability', but (as I will argue) some theorems suffer from the latter kind of problem without suffering from the former. The most common issue, which arises for *each* of the theorems discussed, is that they are *representationally limited* in a number of important respects; that is, they leave us with credence and utility functions which seem fundamentally incapable of modelling the actual credence and utility states of ordinary agents (i.e., an issue of *plausibility*).

The final part of the work seeks to improve the state of characterisational representationism. Chapter 8 develops a new representation theorem aimed at resolving the worst of the *satisfiability*, *plausibility*, *uniqueness* and *circularity* issues found with previous theorems—though it does this at the cost of an essential appeal to unreduced mental notions. I first show how to construct a theorem which is ontologically similar to Jeffrey's, but formally more similar to Ramsey's—but which does not suffer from the infamous problems associated with Ramsey's notion of ethical neutrality, and which has stronger uniqueness results than Jeffrey's theorem. Furthermore, it is argued that the new theorem's preference conditions are descriptively reasonable, even for ordinary agents, and that the credence and utility functions associated with this theorem are capable of a wide range of non-ideal agents—including those who: (i) might have credences and utilities only towards non-specific propositions, (ii) are probabilistically incoherent, (iii) are deductively fallible, and (iv) have distinct credences and utilities towards logically equivalent propositions.

Finally, Chapter 9 is a summary of the thesis, and a look at the present state of characterisational representationism and the naturalisation project.

CHAPTER TWO

Background

The purpose of this chapter is to supply the terminological and conceptual background that will be needed for the rest of the work. §2.1 focuses on the notions of *credence* and *utility*, and their numerical representation, while §2.2 takes a closer look at the concept of *preference*. Then, in §2.3, I outline a very simple representation theorem for the measurement of hardness and clarify the most basic notions (weak orderings, *T*-representation, uniqueness) involved in the statement of representation theorems in general. In §2.4, I look at decision-theoretic representation theorems in particular, outlining the key features of a typical classical expected utility (CEU) theorem and distinguishing them from non-classical utility (NCU) theorems. Finally, in §2.5, I precisify the *Decision-theoretic Interpretation* of a representation theorem, which forms the basis for their philosophical application.

2.1 Credences, utilities, and the representation thereof

I will assume, without argument, a *minimal realism* about graded propositional attitudes; that is, ordinary agents in ordinary circumstances have, as an objective matter of fact, credences and utilities. Moreover, I will assume (pace Harman 1986, and Holton forthcoming) that credence talk is not a mere *façon de parler* for talk about outright beliefs, and likewise for utilities and desires, *mutatis mutandis*.

Let us be clear on what this means. In all that follows, I will use ‘credences’ and ‘utilities’ to refer to the graded propositional attitudes that are the main subject of this work. It will also be helpful to distinguish two different senses in which beliefs can be graded. Consider the following ordinary language locutions:

(1) John is *certain* that his fear of leprechauns won’t get the best of him this time.

(2) Frank is *unsure* whether he is in the matrix.

(3) I am *25% certain* that I will have paid employment next year.

(4) I am *more confident* that I have made a mistake somewhere than I am in the validity of this proof.

(5) Bob is *much more certain* that Jack stole his cake than that Jill did.

These are all attributions of an attitude to a thinking subject; each refers to a kind of *credence state*. Examples (1) to (3) attribute what we might call *absolute credence states*, while (4) and (5) are attributions of a *relative credence state* (also sometimes called *comparative beliefs* or *qualitative probabilities*). A similar relative/absolute distinction exists between graded desire states. However, ordinary language already has a term for relative desirabilities—namely, 'preference' (in the *mentalistic* sense, to be discussed below). For this reason, 'utilities' will always refer to the absolute states.

Ordinary language attributions of absolute credence states ascribe to an agent at a time an opinion regarding a proposition which comes with a particular level of confidence, where these different levels are usually marked out using one of a variety of terms, including 'certain', 'almost positive', 'fairly sure', 'unconfident', and so on. Examples (1) to (3), and countless others, suggest that different absolute credence states can be individuated *via* two factors: the *proposition* that the state is about, and the particular *level* of confidence that attaches to it. Thus, for example, *being certain that P* is a distinct state from *being unsure whether P*, as they involve different levels of confidence; and both of these states are distinct from *being certain that Q*, for distinct propositional relata *P* and *Q*. Similar points can be made with respect to utilities, *mutatis mutandis*.

Relative credences are the kinds of states one might attribute through such phrases as 'I am *more confident* that I have made a mistake somewhere than I am in the validity of this proof'. Instead of attributing an absolute level of confidence to an agent, relative credence attributions ascribe a somewhat different kind of attitude—that of finding a given proposition more, less, or equally likely to another proposition. Examples like (5) also suggest that relative credence attributions can be used to convey not just *ordinal* information, but also information about relative *strengths* with which propositions are believed.

The ubiquity of ordinary language attributions like (1) to (5) indicates that credences and utilities are not merely high-level theoretical constructs whose function is to relate and explain behavioural patterns, with no deep connection to any notions in folk psychology and everyday attitude attributions. At most, we might say that, in academic contexts, *credence* and *utility* are semi-technical notions grounded thoroughly in the folk conception of the mind. There is of course room for the stipulation and development of technical notions in the psychological sciences, but those will not be of interest to us here. The kinds of questions which we will focus on in this work relate to the kinds of states that the folk refer to when they assert thinks like (1) to (5).

Particularly important for our purposes is the fact that different levels of confidence are frequently represented numerically, as in example (3). In academic disciplines which deal with credences and utilities, mathematical models of *total* credence states—that is, a single agent's full range of absolute and relative credences—usually take the form of numerically-valued functions defined on a set of propositions. More specifically, in most cases the models take the form of a *credence function*:

Definition 2.1: Credence function

f is a *credence function* iff $f\colon \mathcal{P} \mapsto [0, 1]$, where $\mathcal{P}$ is a set of propositions

This definition of a credence function does *not* require that propositions in $\mathcal{P}$ are sets of worlds. For the purposes of Definition 2.1, we need only take propositions to be abstract entities with semantic values that make them fit to serve as the contents of our thoughts.

Philosophers sometimes complain that such numerical models are unrealistic, as ordinary agents "don't have numbers in their heads". This is a misconception: real-world objects don't come with pre-attached numbers describing their weights, lengths, and volumes, but this is no reason to think that ordinary objects lack such quantities. As is the case with physical quantities, all that matters is that our credences have a particular kind of *structure* such that they can be usefully represented with numbers—on this, see §2.3. That our credences *do* have such structure is, of course, a question open for debate—though given the great successes achieved using numerical models of total credence states, it's unlikely that they will go away any time soon.

In what follows, I will use '$\mathcal{B}el$' to refer to any function designed to numerically model a total credence state, while '$\mathcal{D}es$' will refer to a numerical model of a total utility state. $\mathcal{B}el$ will usually be a credence function; for exceptions, see §8.3.3 and Appendix B. If $\mathcal{B}el$ accurately models an agent's total credence state then it will pair each proposition towards which the agent has some credence with a value that appropriately captures the degree of confidence attached to that state for the agent in question. I will leave open exactly what is required for a model to be *accurate* or for it to *appropriately model* agents' credences: it seems unlikely that this notion can be usefully precisified prior to an already-established metaphysics of graded attitudes.

Credence functions are lacking in internal structure. In general, representation theorems will impose more structure upon their credence functions—that is, they will imply that $\mathcal{B}el$ satisfies certain properties. The vast majority of contemporary philosophical discussion has focused on a particular kind of credence function, namely, *probability functions*:

Definition 2.2: Probability function

$f\colon \mathcal{X} \mapsto [0, 1]$ is a *probability function* iff $\mathcal{X}$ is an algebra of sets on some set $\mathcal{Y}$, and:

(i) $f(\mathcal{Y}) = 1$

(ii) For all $x \in \mathcal{X}$, $f(x) \geq 0$

(iii) For all $x, y \in \mathcal{X}$, if $(x \cap y) = \emptyset$, then $f(x \cup y) = f(x) + f(y)$

Furthermore, f is a *countably additive probability function* iff f is a probability function and:

(iv) If $x_1, x_2, x_3, \ldots$ is in $\mathcal{X}$ and x_1, x_2, x_3 are pairwise jointly inconsistent, then $f(x_1 \cup x_2 \cup x_3 \cup \ldots) = f(x_1) + f(x_2) + f(x_3) + \ldots$

Importantly, probability functions are defined on *algebras*:

Definition 2.3: Algebra of sets

$\mathcal{X}$ is an *algebra of sets* on $\mathcal{Y}$ iff, $\mathcal{X}$ is a nonempty set of subsets of $\mathcal{Y}$, and for every $x, y \in \mathcal{X}$,

(i) $\mathcal{X} \backslash x \in \mathcal{X}$

(ii) $(x \cup y) \in \mathcal{X}$

Furthermore, $\mathcal{X}$ is a *σ-algebra* iff it is an algebra of sets on $\mathcal{Y}$ and:

(iii) If $x_1, x_2, x_3, \ldots$ is in $\mathcal{X}$, then so is $x' = (x_1 \cup x_2 \cup x_3 \cup \ldots)$

Finally, an algebra $\mathcal{X}$ is *bottomless* just in case:

(iv) For each $x \in \mathcal{X}$, there are two non-empty $y, y' \in \mathcal{X}$ such that $(y \cap y') = \emptyset$, and $(y \cup y') = x$

It follows from conditions (i), (ii), and the fact that $\mathcal{X}$ is non-empty that $\mathcal{Y}$ and the empty set $\emptyset$ are both in $\mathcal{X}$.

It is easy to see that every probability function on a set $\mathcal{P}$ of propositions (usually understood as an algebra on a set of worlds $\mathcal{W}$) is also a credence function, but not *vice versa*. A credence function need not satisfy any of (i) to (iii), and the domain of a credence function need not be an algebra. In all that follows, I will reserve the phrase 'probability function' for functions which satisfy Definition 2.2. Similarly, 'probabilities' will *only* be used to refer to the values of a probability function.[8] If an agent's total credence state is accurately modelled by a probability function—or, more specifically, a probability function defined on an algebra constructed from a set of *possibilities*—then we can say that the agent is *probabilistically coherent*. (As I will discuss further in Chapter 4, while every probabilistically coherent agent's credences can be modelled by a probability function, not every probability function must model a probabilistically coherent agent's credences.)

There may, however, be probabilistically incoherent agents; or, in another turn of phrase, non-probabilistic credences. Many NCU representation theorems involve non-probabilistic credence functions, such as Choquet capacities:

Definition 2.4: Choquet capacity

$f : \mathcal{X} \mapsto [0, 1]$ is a *Choquet capacity* iff $\mathcal{X}$ is an algebra of sets on some set $\mathcal{Y}$, and:

(i) $f(\mathcal{Y}) = 1$

(ii) $f(\emptyset) = 0$

(iii) For all $x, y \in \mathcal{X}$, if $x \subset y$ then $f(y) \geq f(x)$

[8] In some cases I will use the word 'likelihood'. Such uses should be understood in its colloquial sense (akin to 'subjective probability'), rather than its technical meaning in probability theory—i.e., where $\mathcal{P}r$ is a probability function, the likelihood of H with respect to E is $\mathcal{P}r(E|H)$.

Capacities are important for the theorems of Tversky and Kahneman (1992) and Schmeidler (1989), amongst many others.

Choquet capacities were introduced as a generalisation of probability functions—a probability function is simply a capacity satisfying a further condition (additivity). The notion of a credence function in a certain respect takes the generalisation several steps further. In the literature on the representation of credence systems, we also find the use of *Dempster-Shafer belief functions* and *plausibility functions* (Shafer 1976, Dempster 1968), and *possibility measures* (Dubois and Prade 1988). Like capacities and probability functions, these can all be taken to be varieties of credence function, distinguished from one another by their structural characteristics.

Importantly, however, characterisational representationism is not committed to representing credences by means of a credence function as defined above. Inasmuch as credence functions have been the focus of discussion, it is primarily due to the scarcity of decision-theoretic representation theorems which represent credence states by any other means. There are, however, strong reasons to look beyond credence functions for the representation of our credences, which always assign a *precise* real value as a measure of credence (see Levi 1974, Kyburg 1992, Hájek and Smithson 2012). Some representation theorems exist which generalise the notion of a credence function still further. For instance, in Alon and Schmeidler's (2014) recent theorem, $\mathcal{B}el$ is an *interval-valued* function; i.e., a function from a set of propositions into a set of *intervals* constrained by [0, 1]. Real-valued credence functions can be taken as a special case of interval-valued functions, in the obvious way.

Total credence states have also been represented by so-called *ranking functions* (see Spohn 1988, 1990), *plausibility measures* (Halpern 2005), and several other kinds of functions which are not (or need not be) credence functions (for an overview of the alternatives, see Huber and Schmidt-Petri 2009, Halpern 2005). There are more possibilities here than we can consider in the available space, so most of my attention will be directed towards credence functions.

Finally, the function $\mathcal{D}es$, designed to represent a total utility state, will always be a *utility function*, which can be more variable in character than credence functions. Usually, a utility function is any function from a non-empty set into the set of real numbers intended to represent an agent's total utility state. However, for different purposes it may be helpful to take the range of a utility function as including infinite cardinals along with the real numbers, or perhaps even intervals of numbers.

2.2 Two kinds of preference

Our focus is on those representation theorems which have been developed for decision theory, which purport to show how a suitably rational system of *preferences* can be represented (in a sense to be precisified shortly). Exactly how the term 'preference' is to be understood varies from one context to another, and we will see various ways of understanding this notion and the objects of preference over the course of this work. Very roughly, though, we can distinguish two broad senses, which we ought to look at before moving on.

The first sense might be called the *mentalistic* understanding of preference, where a *mentalistic preference* for P over Q is understood as the mental state of finding P more desirable than Q.[9] The objects of mentalistic preference tend to be understood as propositions, though there may also be thought to exist primitive objectual preferences as well. It is this sense of 'preference' that appears to be what Richard Jeffrey had in mind when he wrote that:

> To say that [P] is ranked higher than [Q] [in the agent's preference ranking] means that the agent would welcome the news that [P] is true more than he would the news that [Q] is true: [P] would be better news than [Q]. (1990, 82)

In the second sense, an agent's preferences are understood as behavioural-dispositional states; hence they might be called *behavioural preferences*. In particular, behavioural preferences are a kind of *choice disposition*—roughly, *S behaviourally prefers x* over *y* just in case, were *x* and *y* her only options, she would choose *x*. This is often described as the *standard* or *orthodox* conception of preference within economics and in many other fields (including philosophy) where decision theory is applied, and is most closely associated with Savage's and similar theorems (see Chapter 5).

The variation in how 'preference' is understood is manifest in the great degree of variation in how the objects of preference are formalised within different decision-theoretic representation theorems. Some theorems will treat preferences as being defined on a set of potential *objects of choice* (usually *bets*, *gambles*, or *acts*), while others will define them on a set of *propositions* (which are not in all cases the kinds of things an agent can choose between).

[9] The terminology being used here is borrowed, with slight modifications, from (Dietrich and List forthcoming). Sobel (1997) refers to the mentalistic sense as *preferences tout court*, and argues that it is the more common, folk understanding of the term. However, in many circles—particularly economics—there is a strong tendency to take 'preference' and 'choice' as more or less synonymous. This is largely due to the influence of revealed preference theory. Some authors are careful to distinguish what I have called mentalistic preferences from choice dispositions, but think that the former are directly manifest in the latter.

On all ways of cashing out the notion, though, preferences are *ternary* relations between an *agent at time* and two *objects of preference* (whatever those objects may be). As we only ever consider a single agent's preferences at a time, each agent's preferences (in whatever sense) are, in the majority of cases, formally modelled using a single binary relation, the *weak preference relation* $\succcurlyeq$:

Definition 2.5: Weak preference

For any two *objects of preference* x and y, $x \succcurlyeq y$ (relative to an agent S) iff S either prefers x to y, or is indifferent between x and y

We can define $\succ$ (strict preference) and $\sim$ (indifference) in terms of $\succcurlyeq$.[10] In particular, we can say that $x \sim y$ iff $x \succcurlyeq y$ and $y \succcurlyeq x$, and $x \succ y$ iff $x \succcurlyeq y$ and $\neg(y \succcurlyeq x)$. For the rest of this work, all $\succcurlyeq$, $\succ$, and $\sim$ (i.e., without superscripts) I will refer to as *preference relations*. Preference relations are always defined on a non-empty set $\mathcal{BOP}$ of *basic objects of preference*.

It is worth saying a few more words about the behavioural conception of preference. The historical basis for the behavioural construal of $\succcurlyeq$ traces back at least to revealed preference theory, as founded by Paul Samuelson, who wrote that "the individual guinea-pig, by his market behaviour, reveals his preference pattern—if there is such a consistent pattern" (1948, 243). Samuelson's project was thoroughly behaviouristic, aimed at "freeing" economics from "any vestigial traces of the utility concept" (Samuelson 1938, 71); i.e., by showing that statements about (mentalistic) preferences and utilities can be recast in terms of choice behaviour.

Even amongst those who might otherwise reject behaviourism, there is still the strong tendency to interpret preference relations as they are found within a standard decision-theoretic representation theorem in behavioural terms. It is *routine* for descriptive decision theorists to describe their theorems' preference conditions as *behavioural* conditions. Examples here are legion, though a particularly telling recent example is a paper entitled 'A Simple Behavioral Characterisation of Subjective Expected Utility' (Blavatskyy 2013), which claims to present "a new behavioral characterization (preference axiomatization) of subjective expected utility" within a Savage-style framework.[11] Likewise, in their recent paper against characterisational representationism, Christopher Meacham and Jonathan Weisberg assume that the typical representation theorem's preference conditions can be taken to "encode [an agent's] behavioural dispositions" (2011, 643).

[10] Sometimes—particularly when $\succcurlyeq$ is allowed to be incomplete—theorists will take $\succ$ and $\sim$ as primitives, with $\succcurlyeq$ being defined in terms of them rather than *vice versa*.

[11] Incidentally, as the title makes clear, this paper is also putting forward an instance of characterisational representationism.

There are, however, well-known problems with the choice-based interpretation of preference relations. I will only briefly discuss two of these; my intention in this work is to cast doubt upon the theorems which best fit with a behavioural conception of preference by highlighting issues which are independent of whether that conception is viable (or can be made viable).[12]

The first problem is that the foregoing (and admittedly rough) characterisation of behavioural preferences is incapable of distinguishing "genuine" (i.e., mentalistic) preferences from indifference.[13] Suppose first of all that *S* is rational, and always chooses the alternative which she prefers. However, even given *S*'s rationality, her choice of *x* instead of *y* (when only *x* and *y* are available) may be the result of a preference for *x*—but it may also be that *S* is indifferent between the two options, and (in the nearest possible world where *x* and *y* are her only options) chose *x* at random because she had to choose one.

This well-known problem has leads naturally to the following refined definition:

> *S behaviourally prefers x* over *y* iff, in situations where there no other options available, *S* is disposed to choose *x*; *S* is *indifferent* between *x* and *y* iff *S* has no dispositions either way

The refinement helps (randomly choosing *x* is not the same thing as being disposed to choose *x*), but perhaps it does not go far enough. Consider the following case, which originates with (Maher 1993, 12-15). Sally is presented with three essentially identical opaque boxes, labelled *x*, *y*, and *z*, and allowed to take one. Suppose that Sally feels no particular desire for *x*, *y*, or *z* over any of the others. However, due to a general sense of angst towards indecision—as a child, she was told horror stories about Buridan's ass—Sally has cultivated a disposition to choose any box labelled *x* in this kind of situation. In this kind of case, Maher argues, Sally's disposition to choose *x* over *y* does not reflect any genuine *feeling* of preference—she is indifferent between all the options—but according to the revised definition, $x \succ y$.

Now, to be sure, a proponent of the behavioural interpretation of $\succcurlyeq$ need not be interested in whether *S feels* a stronger desire for *x* over *y*. Perhaps the intended interpretation of $\succcurlyeq$ is not supposed to capture perfectly what the folk mean by 'preference', but instead a technical notion which should be divorced from the introspectively accessible intensities of desire that we feel towards objects of choice (cf. Ramsey 1931, 171-2). We cannot reject the behavioural conception of preference just because it's not coextensive with the mentalistic conception.

The second issue with the behavioural conception of preference seems to me the more serious, however, and concerns the counterfactuals involved. In particular, the problem is that the nearest possible worlds in which the antecedents of the counterfactuals in the

[12] See especially §5.2–4, and §6.1.

[13] See also Joyce (1999, 19-22, 99-102).

definition are true might be very far off indeed. We need to consider a scenario, for each pair of possible objects of choice that are presently available to the agent, in which *just* those two options are on the table, so to speak. For almost all decision situations, there are a vast number of different possible options to choose from. For instance, where *x* and *y* are two arbitrary *acts*, it's hard to even imagine what a world must be like for *only x* and *y* to be available, if indeed there are any such worlds at all. Certainly, if they even exist, these are worlds far different than the one in which the decision-maker is actually making any decisions—and they are likely to be worlds where her credences and utilities are quite different than they are in the actual world. It is hard to imagine why an agent's dispositions in such circumstances should be very closely related to what might be going on inside her head here in the actual world.

In what follows, I will assume that—despite these issues—a behavioural definition of $\succcurlyeq$ can be made viable, and has roughly the form that it was presented with above. As I will argue below, there are more troubling concerns for characterisational representationism, if it appeals to a representation theorem designed around the behaviouristic notion of preference.

2.3 The representational theory of measurement

In order to understand the thesis of characterisational representationism, it will be helpful to have a clear idea of what representation theorems consist in. I will begin with a very simple example of a representation theorem which does not originate from decision theory. Suppose we have a set of 1000 concrete objects, $\mathcal{OB} = \{ob_1, ob_2, \ldots, ob_{1000}\}$, where some of these objects may be *just as hard as* others, while some may be *harder than* others. (Assume for simplicity that each object has a uniform hardness.) Our goal is to find a way to formally represent this quantity of hardness, in a sense to be made precise shortly.

Let $\succ^h$ stand for the *harder than* relation, and $\sim^h$ the *just as hard as* relation. These two relations form two non-overlapping parts of the *at least as hard as* relation $\succcurlyeq^h$; so, for all ob_i and ob_j,

$$ob_i \succcurlyeq^h ob_j \text{ iff either } ob_i \sim^h ob_j \text{ or } ob_i \succ^h ob_j$$

We suppose that in at least one direction, $\succcurlyeq^h$ holds between every pair of objects—that is, for all ob_i and ob_j,

$$ob_i \succcurlyeq^h ob_j \text{ or } ob_j \succcurlyeq^h ob_i$$

In this case we say that $\succcurlyeq^h$ is *complete* (on $\mathcal{OB}$). Furthermore, it is very plausible that $\succcurlyeq^h$ is *transitive*; that is, for all ob_i and ob_j,

> If $ob_i \succcurlyeq^h ob_j$ and $ob_j \succcurlyeq^h ob_k$, then $ob_i \succcurlyeq^h ob_k$

The satisfaction of these two conditions, *transitivity* and *completeness*, means that $\succcurlyeq^h$ on $\mathcal{OB}$ is a *weak ordering*.

> **Definition 2.6: Weak ordering**
> A binary relation $\succcurlyeq^x$ is a weak ordering iff $\succcurlyeq^x$ is transitive and complete

In an intuitive sense, $\succcurlyeq^h$ orders $\mathcal{OB}$ into a sequence of groups according to their hardness. The first part of $\succcurlyeq^h$, the indifference relation $\sim^h$, is *symmetric* (i.e., $ob_i \sim^h ob_j$ implies $ob_j \sim^h ob_i$) and transitive. Because of this, $\sim^h$ can be understood as sorting the objects into groups with exactly the same degree of hardness. The second part of $\succcurlyeq^h$, $\succ^h$, which is *antisymmetric* (i.e., $ob_i \succ^h ob_j$ implies $\neg(ob_j \succ^h ob_i)$) and transitive, can then be understood as ordering those groups into a sequence from the most to the least hard.

We wish to represent this weak ordering numerically, in the sense of assigning numbers to the objects to represent their place in the order, with larger numbers being used to represent greater degrees of hardness. For this we appeal to a *representation theorem*. It turns out that, given our suppositions about $\succcurlyeq^h$, we can prove the existence of a function f which assigns a natural number to each object in $\mathcal{OB}$ such that for all such objects ob_i and ob_j,

> $ob_i \succcurlyeq^h ob_j$ iff $f(ob_i) \geq f(ob_j)$

In the jargon, *f represents $\succcurlyeq^h$ on $\mathcal{OB}$*. Note, though, that this is a highly technical usage of the term ‘represents’, and it will be helpful for what follows to distinguish this technical usage of ‘represents’ from the everyday, folk conception of representation. Let us use ‘*T*-represents’ for the technical notion:

> **Definition 2.7: *T*-representation of a binary relation**
> A function $f\colon \mathcal{X} \mapsto \mathbb{R}$ *T-represents* a binary relation $\succcurlyeq^x$ on $\mathcal{X}$ iff, for all $x, y \in \mathcal{X}$, $x \succcurlyeq^x y$ iff $f(x) \geq f(y)$

So defined, T-representation is relation which holds between a function f and a relation $\succcurlyeq^{x}$ on a set $\mathcal{X}$ iff f is a mapping from $\mathcal{X}$ into $\mathbb{R}$ that preserves the structure of $\succcurlyeq^{x}$ on $\mathcal{X}$.[14] With only a slight abuse of this technical usage, we might also say that for all $x \in \mathcal{X}$, $f(x)$ T-represents x whenever f T-represents $\succcurlyeq^{x}$ on $\mathcal{X}$.

In an important sense, what is required for one thing to T-represent another is far more demanding than we would expect given the ordinary notion of representation, which does not require such strict correspondence of structure. In the ordinary sense, a portrait might represent a famous figure, and indeed it might do so quite well (or quite poorly), without resembling the figure perfectly. The T-representation relation shown to exist by a representation theorem more closely resembles an infinitely high-definition photograph than it does a portrait. The required precision also means that necessary and sufficient conditions for T-representation are demanding. That one system cannot be T-represented by another system should not be taken to imply that the latter cannot adequately represent the former according to the ordinary notion of representation.

In the most general sense, a representation theorem is a (mathematically provable) statement to the effect that if certain conditions are satisfied, then there exists a structure-preserving mapping between two previously defined kinds of sets—typically, a set of concrete objects all sharing a quantitative property to differing degrees, and a set of numbers. The relevant structure to be preserved can be specified by means of a *relational system*; that is, a sequence of the form $<\mathcal{X}, R_1, \ldots, R_n>$, where $\mathcal{X}$ is a non-empty set and $R_1, \ldots, R_n$ are relations defined on $\mathcal{X}$.[15] The intuitive idea is that the relations $R_1, \ldots, R_n$ characterise the *relational structure* of the empirical domain $\mathcal{X}$—or at least the relational structure that we are interested in capturing numerically.

Say that $<\mathcal{X}, \succcurlyeq^{x}>$ is a *finite weak order* iff $\mathcal{X}$ is non-empty but finite and $\succcurlyeq^{x}$ on $\mathcal{X}$ is an arbitrary weak ordering. The theorem adverted to just above would then be:

Theorem 2.1: Simple finite ordinal scale
If $<\mathcal{X}, \succcurlyeq^{x}>$ is a finite weak order, then there exists a function $f\colon \mathcal{X} \mapsto \mathbb{R}$ that T-represents $\succcurlyeq^{x}$ on $\mathcal{X}$

In effect, Theorem 2.1 asserts that any finite weak order $<\mathcal{X}, \succcurlyeq^{x}>$ is isomorphic to some numerical relational system $<\mathbb{R}^*, \geq>$, where $\mathbb{R}^* \subseteq \mathbb{R}$. Given our assumptions, the *hardness relational system* $<\mathcal{OB}, \succcurlyeq^{h}>$ is a finite weak order, and so can be given a simple ordinal T-representation: one can *precisely* replicate the structure of $\succcurlyeq^{h}$ on $\mathcal{OB}$ using $\geq$ on some set of numbers $\mathbb{R}^*$.

[14] Usually isomorphic mappings are desired, but other kinds of structure-preserving mappings are countenanced in the representational theory of measurement. See (Swoyer 1991) for the *minimal* sense in which $\mathcal{F}$ must be structure-preserving.

[15] Formally, n-ary relations are modelled set-theoretically as ordered n-tuples.

Note, however, that Theorem 2.1 says nothing about hardness *directly*. To see this, note that while $\succcurlyeq^{x}$ could symbolise the *at least as hard as* relation, it could also symbolise *to the left of*, or *at most as funny as*, or any other binary relation whatsoever. The application of the theorem to the measurement of hardness depends on an *interpretation*—the substitution of a purely formal system $<\mathcal{X}, \succcurlyeq^{x}>$ for a particular system $<\mathcal{OB}, \succcurlyeq^{h}>$ with the adequate structure.

There is clear value in constructing a numerical *T*-representation. Theorem 2.1 shows that the hardness relational system has the same structure as a set of numbers weakly ordered by the *greater than* relation. By virtue of this similar structure, we can engage in what Swoyer (1991) helpfully refers to as *surrogative reasoning*, or reasoning using the numerical system so as to draw conclusions about the empirical (and *non-numerical*) system that it *T*-represents. We are very adept at recognising quickly when one number is greater than, less than, or equal to another number, so to label one object ob_1 with a number n and another object ob_2 with m supplies us with an immediately accessible and easily manipulable system with which to reason about the relative hardnesses of ob_1 and ob_2.

The theorem also serves to highlight exactly which relations between the natural numbers can be used to (validly) reason surrogatively about empirically interesting relations between the objects. In our example, $\succcurlyeq^{h}$ is *T*-represented by f in the form of $\geq$, and given our assumptions, this implies that:

$$ob_i \sim^{h} ob_j \text{ iff } f(ob_i) = f(ob_j)$$

And:

$$ob_i \succ^{h} ob_j \text{ iff } f(ob_i) > f(ob_j)$$

However, other possible mathematical relationships between the numbers $f(ob_i)$ and $f(ob_j)$ need not correspond to any interesting relationship that holds between the objects in $\mathcal{OB}$. For instance, suppose that $f(ob_i) = 2.f(ob_j)$. It would be mistake to infer that ob_i will be *twice as hard as* ob_j, because the *twice as hard as* relation is nowhere specified in the relational system $<\mathcal{OB}, \succcurlyeq^{h}>$. An equivalent way to make this point is to note that there are *infinitely* many ways to assign natural numbers to the objects in $\mathcal{OB}$ so as to accurately preserve their places within the $\succcurlyeq^{h}$ order, and the fact that $f(ob_i) = 2.f(ob_j)$ on one assignment of values f does not imply that $f^*(ob_i) = 2.f^*(ob_j)$ on any other assignment f^*. In the jargon, we would say in this circumstance that f is *unique up to monotone transformation*, where a monotone transformation $\mathcal{T}$ is a function that assigns new values such that:

$$f(ob_i) \geq f(ob_j) \text{ iff } \mathcal{T}(f(ob_i)) \geq \mathcal{T}(f(ob_j))$$

It is only the mathematical information which is common to *all* of these assignments (i.e., their ≥-order) that is empirically meaningful and available for surrogative reasoning; anything else is an artefact of the particular numerical assignment arbitrarily chosen from an infinite set of equally valid measures. To *T*-represent any further information, such as *ratios of hardness*, we would need to use a more structured relational system than $<\mathcal{OB}, \succcurlyeq^{h}>$—and we would also need much more demanding conditions to establish the existence of an appropriate *T*-representation.

2.4 Decision-theoretic representation theorems

The simple relational system $<\mathcal{BOP}, \succcurlyeq>$ captures the structure of an agent's *basic system of preferences* at a particular time. The aim of a *decision-theoretic* representation theorem is then to develop a suitable, and reasonably unique, numerical *T*-representation of $\succcurlyeq$ on $\mathcal{BOP}$. Unless otherwise specified, all uses of 'representation theorem' will henceforth refer only to *decision-theoretic* representation theorems.

The vast majority of contemporary theories of decision-making treat an agent's basic system of preferences as being determined by the interaction of (at least) two distinct measurable factors—her credences and her utilities. Correspondingly, the aim of these theorems is to show that an agent's basic system of preferences can be *T*-represented by a single numerical ordering determined in turn by the combination of (at least) two functions corresponding to the agent's credences and utilities: $\mathcal{B}el$ and $\mathcal{D}es$.

There are two basic kinds of theorem we will look at: *classical expected utility* (CEU) theorems and *non-classical utility* (NCU) theorems. CEU theorems are more widely known and discussed by philosophers; they are often taken to form the foundations for orthodox Bayesian approaches to rational decision-making. NCU theorems are more generally favoured within psychology as a descriptive enterprise aimed at characterising the actual decision-making behaviour of ordinary agents. In what follows, I will first give an outline of the main features of a typical CEU theorem, before distinguishing CEU from NCU theorems.

The standard model of a decision situation takes the form of a matrix:

States **Options**	s_1	s_2	s_3	…
x	o_2	o_1	o_2	…
y	o_1	o_2	o_2	…
…	…	…	…	…

We find in this model several key elements:

- A number of possible *states* (or *possibilities*, *ways the world might be*, *events*, etc.). These should be pairwise inconsistent and jointly exhaustive of the possibilities (or at least the possibilities the decision-maker has some credence in).
- A number of *options* (e.g., *acts*, *decisions*, *gambles*, etc.). These are the items we aim to decide between, the basic objects of preference. They will typically have different outcomes under different possibilities. These should constructed such that choice of one option precludes the choice of any other.
- A number of possible *outcomes*, or the consequences of choosing a particular option given a particular state. These need not partition the space of possibilities, but they should be mutually exclusive.

The purpose of the decision matrix is to determine a *preference ranking* on the options according to some decision-making principle. According to CEU, that principle is *expected utility maximisation*: states are assigned credences, outcomes are assigned utilities, and the preferred act should have the highest credence-weighted average for its associated outcomes. CEU also imposes the requirement that credences ought to be probabilities—or, in another manner of speaking, the theory is only applicable to probabilistically coherent agents.

It is worth noting that here and below I will use ‘expected utility maximisation’ in a relatively loose way: an agent maximises expected utility just in case she chooses the option with the highest *credence*-weighted average utility, regardless of whether those credences are probabilities. In the mathematical jargon, ‘expectation’ is *defined* in terms of probability functions: the *expected value* (*EV*) of a numerically-valued function f in a single discrete variable x is:

$$\sum_x f(x).\mathcal{P}r(x)$$

where $\mathcal{P}r$ is a pre-specified probability function. However, contemporary theorists usually have a more general notion of ‘expectation’ in mind when they speak of, for example, Choquet *expected* utility theory—according to which preferences can be represented using the basic form:

$$\sum_x \mathcal{D}es(f(x)).\mathcal{B}el(x)$$

where $\mathcal{B}el$ need only be a capacity. In what follows, then, my use of ‘$\mathcal{EU}$’ will designate *expected utility functions*, in the looser sense of ‘expected utility’. The defining characteristic of any $\mathcal{EU}$ function is that assigns a numerical value n to a *basic object of preference*, where n is equal to the *credence-weighted average utility* of the possible *outcomes* associated with that object (where the credences in question *need not be* probabilities).

As an example of how CEU works, suppose that we fill in the values of a decision matrix like so:

States **Options**	$s_1 = 0.25$	$s_2 = 0.25$	$s_3 = 0.5$
x	$o_2 = 2$	$o_1 = 1$	$o_2 = 2$
y	$o_1 = 1$	$o_2 = 2$	$o_2 = 2$
z	$o_3 = 3$	$o_3 = 3$	$o_2 = 2$

The states, $s_1 – s_3$, we assume are mutually exclusive and jointly exhaustive, and so their probabilities sum to 1. The expected utility of x is:

$$\mathcal{E}\mathcal{U}(x) = 0.25(2) + 0.25(1) + 0.5(2) = 1.75$$

This is equal to the expected utility of y:

$$\mathcal{E}\mathcal{U}(y) = 0.25(1) + 0.25(2) + 0.5(2) = 1.75$$

According to CEU, then, $x \sim y$. However, the expected utility of z is 2.5:

$$\mathcal{E}(z) = 0.25(3) + 0.25(3) + 0.5(2) = 2.5$$

The final ranking we arrive at is thus $z \succ x \sim y$. In this way, each option in a decision situation can be assigned a numerical value according to its position in the preference order, with higher expected utility values sitting higher in the order.

Representation theorems for decision theory come in all shapes and sizes; however, every such theorem (for either CEU or NCU) will formalise the basic elements of the standard decision matrix in one way or another. It is impossible to state in a general fashion how this is done: different theorems may employ different primitives, different constructions out of those primitives, or require different conditions on preferences, and they may lead to very different T-representations with varying degrees of uniqueness. Following Savage's (1954) seminal contribution, however, most representation theorems are based around three sets of entities—a set of *outcomes*, a set of *possibilities* (or *states*), and a set of *acts* (formally modelled as functions from states to outcomes)—with $\succcurlyeq$ being defined in the first instance on the set of acts. We will look Savage's formal system in much greater detail in Chapter 5; for now, the specifics can be set aside.

With the basic formal elements specified, we find a statement of a number of *preference conditions*, which we will label C, such as the requirement that $\succcurlyeq$ on $\mathcal{BOP}$ is a weak ordering.[16] The typical CEU theorem then has the following general form:

Typical CEU Theorem
If $\succcurlyeq$ on $\mathcal{BOP}$ satisfies the stated conditions C, then there exists a function $\mathcal{EU}$: $\mathcal{BOP} \mapsto \mathbb{R}$ that T-represents $\succcurlyeq$ on $\mathcal{BOP}$; i.e., for all $x, y \in \mathcal{BOP}$,

(i) $x \succcurlyeq y$ iff $\mathcal{EU}(x) \geq \mathcal{EU}(y)$,

where $\mathcal{EU}$ is an expected utility function determined by functions $\mathcal{B}el$ and $\mathcal{D}es$ which satisfy properties R (esp., $\mathcal{B}el$ must be a probability function)

Some of the conditions specified in C may be *necessary* for the existence of the T-representation, where this means that their satisfaction is implied by the assumption that the relevant T-representation exists. In the event that the *all* of the conditions C are necessary, the theorem can be stated as a biconditional instead of taking the conditional form given here. However, in almost all cases there will also be a number of non-necessary conditions included in C as well; these typically take the form of an existential condition. It is extremely difficult to discover conditions that are both jointly sufficient and individually necessary for the existence of an $\mathcal{EU}$ T-representation.

Uniqueness conditions are usually also given alongside the statement of a representation theorem. What I will call the *Standard Uniqueness Condition* is quite strong:

Standard Uniqueness Condition
$\mathcal{B}el$ is unique and $\mathcal{D}es$ is unique up to positive linear transformation (i.e., unique up to multiplication by a positive real number and the addition of a constant)[17]

It is important to be clear on the sense in which $\mathcal{B}el$ and $\mathcal{D}es$ are unique. The Typical CEU Theorem requires that $\mathcal{EU}$ is an expected utility function determined by some $\mathcal{B}el$ and

[16] A theorem may also require a number of *purely structural conditions* (i.e., conditions that do not refer to the preference relation), which lay down any restrictions or assumptions that are supposed to hold for the sets involved in the statement of the theorem. For instance, a purely structural axiom might specify that the set of states is finite or uncountably infinite, or that the set of outcomes is finite. Sometimes, purely structural axioms are left implicit, or built into one of the definitions that the theorem employs.

[17] Most decision theorists assume that utilities are only measurable on an interval scale, with no sense to be made of an *absolute zero* utility state: utilities can have different *strengths*, and one outcome may be much more or much less desirable than another, but we cannot say (for example) that a given outcome is twice as desirable as another. (Compare the measurement of temperature: 40° C is much warmer than 20° C, but it is not *twice* as warm as the scale depends on an arbitrary choice of unit and zero point.) If this is correct, then we might say that $\mathcal{D}es$ is *effectively* unique under the Standard Uniqueness Condition; that is, differences between positive linear transformations of a $\mathcal{D}es$ function are merely notational.

$\mathcal{D}es$ pair, with *specific properties R, combined in a particular way*. Regarding $\mathcal{B}el$, then, the Standard Uniqueness Condition only says that there is exactly one function *with the relevant properties*, which, *when combined in a particular way with some appropriate* $\mathcal{D}es$, will allow us to T-represent $\succcurlyeq$ on $\boldsymbol{\mathcal{BOP}}$. The relevant properties for $\mathcal{B}el$ will usually include such things as *being defined on a particular set* and *being a probability function*.

The uniqueness condition does *not* imply that the only way to T-represent $\succcurlyeq$ on $\boldsymbol{\mathcal{BOP}}$ is *via* $\mathcal{EU}$-maximisation using some $\mathcal{B}el$ and $\mathcal{D}es$ with properties R. For example, other T-representations might involve either a $\mathcal{B}el$ function without the specified properties (e.g., a *non-probabilistic* function), or they might involve a wholly different combination rule. An analogous point holds for $\mathcal{D}es$: the uniqueness condition only asserts that it's unique up to positive linear transformation *under the condition that* the final form of the T-representation of $\succcurlyeq$ is held fixed. As we will see in §2.5 and §3.2, accounting for a proper interpretation of the uniqueness condition is extremely important for fleshing out the details of characterisational representationism.

The $\mathcal{EU}$ function we arrive at might take a wide variety of forms, dependent on the characteristics of the formal system employed. A simple example of an $\mathcal{EU}$ function would be as follows. Suppose that our basic options are understood to be potential acts the agent might take in the given situation, and a given act is formalised as a function $\mathcal{F}$ from a number n of states to particular outcomes. Suppose also that the final T-representation is such that every state gets a particular probability (assigned by $\mathcal{B}el$) and every outcome gets a particular utility value (assigned by $\mathcal{D}es$). Then we might have $\mathcal{EU}$ determined as:

$$\mathcal{E}(\mathcal{F}) = \sum_i^n \mathcal{B}el(s_\mathrm{i}).\mathcal{D}es(\mathcal{F}(s_\mathrm{i}))$$

This is, roughly, the T-representation of $\succcurlyeq$ arrived at by Savage (see §5.1.2). However, there are many other theorems which involve expected utility maximisation—compare, for instance, Theorem 6.2, Theorem 6.3, and Theorem 8.3.

So much for the typical CEU theorem; we now move on to NCU theorems. These theorems tend to be very similar to CEU theorems in their formal underpinnings. Often, NCU theorems involve exactly the same formal structures as Savage's paradigmatic multiset theorem—albeit with weaker preference conditions, such that a distinct and usually more general style of T-representation is arrived at. Specifically, NCU theorems satisfy at least one of the following:

(a) $\mathcal{B}el$ need not be a probability function

(b) Other functions besides $\mathcal{B}el$ and $\mathcal{D}es$ are employed in the T-representation of $\succcurlyeq$

(c) $\mathcal{B}el$ and $\mathcal{D}es$ jointly T-represent $\succcurlyeq$ according to some combination rule other than expected utility maximisation

Essentially, any representation theorem which does not leave us with a T-representation of agents' preferences as being determined by probabilistic expected utility maximisation is an NCU theorem. The following three examples should be helpful:

* As above, Choquet expected utility models T-represent $\succcurlyeq$ in a manner extremely similar to Savage's CEU theorem, however $\mathcal{B}el$ need only be a capacity.
* Buchak's (2013) T-representation of $\succcurlyeq$ involves a probability function $\mathcal{B}el$, a utility function $\mathcal{D}es$, and a third real-valued function $\mathcal{R}$ that is intended to reflect the degree to which an agent is *risk averse*.
* Alon and Schmeidler's (2014) T-representation involves a $\mathcal{B}el$ which is not a credence function, and combines $\mathcal{B}el$ and $\mathcal{D}es$ according to the so-called *maxmin* rule.

Several important examples of NCU theorems, including Buchak's and Alon and Schmeidler's, are discussed in more detail in Appendix B. Theorem 8.3, developed in Chapter 8, also counts as NCU by virtue of its non-probabilistic $\mathcal{B}el$.

Before we move on to the interpretation of these theorems, it is important to note the specific kind of theorem that I will be focusing on in this work. In particular, I have limited my attention to those theorems which take *preference* relations, and *only* preference relations, as primitive; these we might call *single-primitive* representation theorems. There also exist *dual-primitive* representation theorems. For example, Joyce's (1999) theorem makes use of $\succcurlyeq$ as well as a second primitive binary relation defined on a set of propositions, $\succcurlyeq^{b}$, which is supposed to represent the agent's relative credences. ($\succcurlyeq^{b}$ is often referred to in the literature as a *qualitative probability relation*.) That is,

Definition 2.8: $\succcurlyeq^{b}$
$P \succcurlyeq^{b} Q$ (relative to an agent S) iff S judges P to be at least as likely as Q

Dual-primitive theorems will typically build $\mathcal{B}el$ primarily out of $\succcurlyeq^{b}$, and are for that reason far less useful to preference functionalists hoping to characterise credences in terms of preferences. In what follows, I will not specify that the theorems under discussion are single-primitive.

I have also chosen not to consider in any depth those theorems which take us from a so-called *system of qualitative probability*, $<\boldsymbol{\mathcal{P}}, \succcurlyeq^{b}>$ where $\boldsymbol{\mathcal{P}}$ is an algebra of propositions, to a probability function $\mathcal{P}r$ which T-represents $\succcurlyeq^{b}$ on $\boldsymbol{\mathcal{P}}$. The role of such theorems in the measurement and characterisation of credences is discussed by Koopman (1940), Suppes (1994), Zynda (2000), and Meacham and Weisberg (2011), amongst others. These theorems might be useful for an attempted reduction of absolute credences to relative credences, but the viability of that project is beyond the scope of the present work.

2.5 Interpretations

We must be very careful to distinguish between a representation theorem and its interpretation (cf. Hampton 1994, Hausman 2000, 100-1). Fundamentally, what each representation theorem tells us is that if some relation $\succcurlyeq^{x}$ defined on an appropriately structured set $\mathcal{X}$ satisfies some set of conditions C, then there are some functions $\mathcal{B}el$ and $\mathcal{D}es$ (plus perhaps others) with such-and-such properties that when combined in the right way jointly T-represent $\succcurlyeq^{x}$ on $\mathcal{X}$. What lessons we draw from such results depends on how we interpret $\succcurlyeq^{x}$, $\mathcal{X}$, $\mathcal{B}el$, $\mathcal{D}es$, and the combination rule—and there are countlessly many possibilities here.

For example, in just the same way that the $\succcurlyeq^{x}$ mentioned in Theorem 2.1 need not be interpreted (or interpretable) as the *at least as hard as* relation, the relation $\succcurlyeq$ mentioned in the statement of a representation theorem need not be a preference relation. For instance, suppose that $\succcurlyeq$ is defined on a set of actions and stands for the *involves at least as many poodle interactions* relation. In this case, the theorem may imply that if an agent's available actions are ranked by the number of poodle interactions they involve and that ranking satisfies C, then $\succcurlyeq$ can be T-represented *via* a probability function $\mathcal{B}el$ and a real-valued function $\mathcal{D}es$. Obviously, in this case, there is no reason to think that $\mathcal{B}el$ and $\mathcal{D}es$ correspond to anything psychologically real or interesting: the fact that the poodle ranking might be T-represented in a certain way is hardly more than a mathematical curiosity.

Importantly, a representation theorem may not have any interesting implications for preferences *at all*. The fact that '$\succcurlyeq$' is *called* a preference relation is not enough to ensure that it can reasonably be so interpreted; it is given that name only because that is the interpretation that theorists desire for it to have. However, it may be the case that the formal relata of $\succcurlyeq$ bear no resemblance whatsoever to the kinds of things that we would ordinarily call *objects of preference*—in that case, there would be little sense of interpreting $\succcurlyeq$, $\succ$, and $\sim$ as genuine *preference relations*. As we will see in §5.2 and §6.1, the interpretation of $\succcurlyeq$ as encoding a subject's preferences (in any sense) can sometimes be very strained, given the formal restrictions imposed by the theorem's conditions on $\succcurlyeq$'s basic relata.

Nevertheless, as noted in Chapter 1, representation theorems are usually understood as telling us something about how certain *agents* can be represented *qua* decision-makers. Here, for instance, is a recent (and intentionally informal) gloss of a CEU theorem and its uniqueness condition given by Christensen (2004, 125):

> If an agent's preferences obey constraints *C*, then they can be represented as resulting from some [effectively] unique set of utilities [$\mathcal{Des}$] and probabilistically coherent degrees of belief [$\mathcal{Bel}$] relative to which they maximise expected utility.

We will need to be more precise than this, however. In particular, we will need to generalise the interpretation to accommodate NCU theorems, clarify the relevant sense of 'representation', and (most importantly) clarify exactly what $\mathcal{Bel}$ and $\mathcal{Des}$ can be taken to be representations of.

In the ordinary sense of the term, representation is, or at least can be, very cheap. In particular, representations need not be *accurate*; indeed in some cases a representation can be considered better *because* of its inaccuracies. A caricature, for example, is a kind of representation where exaggeration is a desirable feature. Because a representation in the ordinary sense need not be accurate, there is a trivial sense in which *anyone* can be represented as an expected utility maximiser; likewise, anyone can be represented as an expected utility minimiser, and as having any set of credences and utilities that we like. We did not need a representation theorem to tell us that agents can be represented as following particular decision rules, this much we can know already. If a representation theorem is to be philosophically useful, there must be a tighter sense of 'representation' involved.

What we want is a sense of 'representation' which is weaker than *T*-representation, where a high degree of *accuracy* is assumed but absolute precision is not a success condition. Talk of *models* in the sciences appears to be like this. The billiard ball model of gasses, for example, or the Bohr model of the atom, are simple representations of empirical phenomena, the components of which are supposed to closely (though usually not precisely) correspond to key features of interest in the phenomena being modelled. Importantly, scientists often make use of *mathematical models*—for example, equations designed to represent population growth and predator-prey interactions—which again are deemed useful insofar as they are accurate, despite diverging from the phenomena that they represent in the finer details. Most descriptive decision theorists would count their work as aiming at the development of broadly accurate mathematical models of the ordinary agents' decision-making process. Such models admittedly contain some idealisations and abstractions, but on the whole they are supposed to capture the basic psychological phenomena associated with decision-making.

Not every aspect of a model is designed to have a representational function. A billiard ball is usually made of resin, but the billiard ball model of gases is not meant to imply that gases are usually made of resin. We can thus distinguish between those aspects of a model that are explicitly supposed to play a representational function, and those which don't. This distinction will prove helpful in spelling out the Decision-theoretic Interpretation of a representation theorem. Firstly, call a mathematical model of an agent's decision-making process *paramorphic* iff the model *accurately* captures the facts about the agent's preferences.[18] A paramorphic model of a decision-maker may or may not adequately model the psychological mechanics which give rise to the agent's preferences (it may, for instance, wholly misrepresent the agent's credences and utilities); all that matters is that it produces the *right* pattern of preferences. On the other hand, call a model *homomorphic* iff it not only accurately captures the preference facts, but also captures the agent's actual credences, utilities, and whatever high-level psychological processes are involved in the agent's decision-making procedure. That is, a homomorphic model provides an accurate depiction of the agent's decision-making which gets both the preference patterns and the underlying psychological mechanics right, while a paramorphic model is any model which produces the right preference patterns. Every homomorphic model will, therefore, be a paramorphic model, but only some paramorphic models will be homomorphic.

We now have the resources to precisify the earlier interpretation. I will begin with a specific example—Savage's theorem—before generalising:

> If $\succcurlyeq$ on $\boldsymbol{\mathcal{A}}$ satisfies the stated conditions C, then there exists a function $\mathcal{EU}$ defined on $\boldsymbol{\mathcal{A}}$ that T-represents $\succcurlyeq$ on $\boldsymbol{\mathcal{A}}$, in the sense that for all $\mathcal{F}, \mathcal{G} \in \boldsymbol{\mathcal{A}}$,
>
> (i) $\mathcal{F} \succcurlyeq \mathcal{G}$ iff $\mathcal{E}(\mathcal{F}) \geq \mathcal{EU}(\mathcal{G})$,
>
> where $\mathcal{EU}$ is an expected utility function determined by a probability function $\mathcal{B}el$ (defined on a set of *events* $\boldsymbol{\mathcal{E}}$) and a utility function $\mathcal{D}es$ (defined on a set of outcomes, $\boldsymbol{\mathcal{O}}$); furthermore, there is only one such probability function $\mathcal{B}el$, and $\mathcal{D}es$ is unique up to positive linear transformation

Let us say that an agent is *preference-rational* with respect to a theorem's preference conditions C just in case her preferences satisfy C. (In the sequel, I will not specify the C with respect to which an agent counts as preference-rational unless it's unclear from the context of the discussion.) A precise *Decision-theoretic Interpretation* of Savage's representation result is then as follows:

> There is a mapping Ψ that pairs each preference-rational agent S with a *paramorphic* model of S as an expected utility maximiser (with respect to $\succcurlyeq$ on $\boldsymbol{\mathcal{A}}$) with credences *at least partially* modelled by a probability function $\mathcal{B}el$ on $\boldsymbol{\mathcal{E}}$ and utilities *at least partially* modelled by $\mathcal{D}es$ on $\boldsymbol{\mathcal{O}}$

[18] The 'paramorphic'/'homomorphic' terminology is borrowed, with slight modifications, from Wakker (2010, 9).

It will be helpful to have a name for the mapping ψ, so we will call it a *modelling scheme*. The uniqueness condition can also be stated:

> Any model of S as an expected utility maximiser (with respect to $\succcurlyeq$ on $\mathcal{A}$) involving a probability function $\mathcal{Bel}'$ on $\mathcal{E}$ and a utility function $\mathcal{Des}'$ on $\mathcal{O}$ will be such that $\mathcal{Bel}' = \mathcal{Bel}$ and $\mathcal{Des}'$ is some positive linear transformation of $\mathcal{Des}$

Because $\mathcal{Des}$ is not wholly unique, Savage's theorem actually establishes the existence of a *class* of modelling schemes. Every Ψ in this class will assign $\mathcal{Bel}$ as a model of the agent's credences (or at least a partial model), but will assign different $\mathcal{Des}$ functions for the agent's utilities (each a positive linear transformation of the others).

There are some important things to note about this interpretation. First of all, inasmuch as the representation theorem in question contains *non-necessary* preference conditions—as do the vast majority—it would be a mistake to suppose that *only* preference-rational agents can be paramorphically modelled in the relevant manner. The theorem tells us that preference-rational agents can be modelled in a certain way; it doesn't tell us that preference-*ir*rational subjects cannot *also* be modelled in that way.

The key point to note about the given interpretation, however, is that $\mathcal{Bel}$ and $\mathcal{Des}$ are not assumed to be *complete* models of S's total credence and utility states. Under this interpretation, if $\mathcal{Bel}$ is not defined for some proposition P—i.e., if $P \notin \mathcal{E}$—then the agent is neither represented as having nor lacking any credence in P—the modelling scheme is silent on this matter (and likewise for $\mathcal{Des}$). Since credences towards the propositions outside of $\mathcal{E}$ are not taken to be involved in determining $\succcurlyeq$, they are unconstrained by the representation theorem: any value (or lack of value) may be assigned to them consistent with the model of S as an expected utility maximiser.

It is *consistent* with the given interpretation that $\mathcal{Bel}$ and $\mathcal{Des}$ be treated as *complete* models of an agent's credences and utilities, but this interpretation would only be wise if we have good reason to think that $\mathcal{E}$ and $\mathcal{O}$ contain all of the entities towards which the subject S has credences and utilities respectively. In Savage's system, $\mathcal{O}$ is usually a set of propositions which are maximally specific with respect to what the agent cares about, so it obviously does not contain all objects of utility. Likewise, in Ramsey's system, $\mathcal{O}$ is described as containing only *possible worlds*. A utility function $\mathcal{Des}$ on $\mathcal{O}$ on either Savage's or Ramsey's conception should never be treated as anything more than a partial model of an agent's total utility state.

The more interesting question much of the time is whether the domain of a theorem's $\mathcal{Bel}$ contains all of the propositions towards which S has credences. As we will see (especially in §5.3 and §6.1.2), this seems highly unlikely for a wide class of theorems. The importance of this point for characterisational representationism has not, so far, been noted in the relevant literature. For the most part, it seems to have been presupposed that

$\mathcal{B}el$ (if not $\mathcal{D}es$) ought to be treated as a complete model—if $\mathcal{B}el(P)$ is not defined, then the agent is represented as *lacking* any credences towards P. However, to the extent that we have good reasons to believe that an ordinary agent will have credences (or utilities) towards a class of propositions not covered by a theorem's $\mathcal{B}el$ or $\mathcal{D}es$, we *ipso facto* have good reasons to believe that the theorem cannot tell us *what it is* to have credences (or utilities) towards those propositions.

So much for the proper interpretation of a CEU theorem; let us now state in general form the *Decision-theoretic Interpretation* of an arbitrary representation theorem, whether it be for CEU or NCU:

> **Decision-theoretic Interpretation**
> There is a modelling scheme Ψ that pairs each preference-rational agent S with a *paramorphic* model of S as following some decision rule (with respect to $\succcurlyeq$ on $\mathcal{BOP}$) with credences *at least partially* modelled by a function $\mathcal{B}el$ (with such-and-such properties) and utilities *at least partially* modelled by $\mathcal{D}es$ (with such-and-such properties)

The Decision-theoretic Interpretation tells us that a preference-rational agent can be *paramorphically* modelled in a certain way, and that the model in question is highly accurate with respect to capturing the agent's preferences—or, at least, her preferences over the relevant set of basic objects of preference.

For many philosophers, the interesting question that then arises is whether the model is *merely* paramorphic, or whether it may also be an homomorphic model—that is, whether the agent who is modelled like so *really is* like so with respect to her internal mental states and decision-making processes. The proponent of characterisational representationism will generally want to say that their favourite representation theorem presents us with an homomorphic model, at least under certain idealised conditions—and if this claim can be justified, then we can use the modelling scheme Ψ to help *define* what it is to have credences and utilities.

CHAPTER THREE

Classical Characterisational Representationism

The philosophical relevance of representation theorems has been the subject of some scepticism in the recent literature (e.g., Hampton 1994, Peterson 2004, Meacham and Weisberg 2011, Easwaran 2014, Dogramaci forthcoming). There are two sides to this scepticism, corresponding to the two primary uses to which representation theorems have been put. On the one side, there is scepticism regarding their *characterisational* relevance. Meacham and Weisberg, for example, spend most of their (2011) paper criticising the idea that "representation theorems play a crucial role in characterising the notions of degree of belief and utility, the graded notions of belief and desire that appear in our folk, descriptive and normative theorising" (642)—that is, characterisational representationism.[19] On the other side, there is scepticism regarding the *normative* relevance of representation theorems—that is, scepticism surrounding the idea that they might play a crucial role in grounding norms such as the principle of expected utility maximisation or the thesis of *probabilism* (that agents ought to be probabilistically coherent).

In many cases, the two sides to this scepticism can be found together. They are closely connected for historical reasons, as those who have tried to apply representation theorems for normative purposes have often made crucial appeal to some form of characterisational representationism. For example, Savage's (1954) and (more explicitly) Maher's (1993) *representation theorem arguments*, which purport to establish both probabilism and the norm of expected utility maximisation, rely on versions of characterisational representationism (see also Christensen 1996, 2001). Indeed, because of the importance of characterisational representationism to representation theorem arguments, scepticism regarding the normative relevance of representation theorems often *depends upon* scepticism regarding their characterisational relevance.

[19] My usage of 'characterisational representationism' derives directly from Meacham and Weisberg's introduction of the name, as here quoted—although they explicitly restrict their criticism to those versions of the view based on CEU representation theorems. In practice, their discussion actually seems to target a much stronger position still—that *S* has such-and-such credences and utilities *if and only if* her preferences satisfy (or approximately satisfy, or would satisfy under certain conditions) the preference conditions associated with some CEU representation theorem. As will become clear, this is a position that I reject, for some of the same reasons that they put forward. However, it is also a position which very few authors (if any) have adopted; including, for instance, the obvious targets, e.g., Ramsey, Savage, Maher—see §3.2 for discussion.

An evaluation of the normative relevance of representation theorems is beyond the scope of this work; for my own part, I side with the sceptics on the view that there is no straightforward argumentative route that begins with a representation theorem and ends with an interesting normative thesis like probabilism. The topic of this work concerns the characterisational relevance of representation theorems, and on this front I don't find the common reasons for scepticism compelling—especially in lieu of a better alternative.

This chapter and the next present a defence of characterisational representationism against the sceptics. In this chapter, I will outline—and ultimately reject—a number of simple versions of characterisational representationism (§3.2–3). This will pave the way for more plausible approaches to applying representation theorems in a characterisational capacity (discussed in Chapter 4).

3.1 Measurement and the problem of separability

It will be helpful to begin with the historical background to characterisational representationism. The earliest use of a decision-theoretic representation theorem can be found in Frank Ramsey's 'Truth and Probability' (1931), which we will discuss in Chapter 7. This involved a CEU theorem in particular, which was developed for the purposes of constructing a system for the measurement of credences and utilities. Since the 1950s, Ramsey's ideas about measurement have been taken up and developed substantially by philosophers, psychologists, and economists looking to create similar measurement procedures (see especially Davidson, Suppes *et al.* 1957, Krantz, Luce *et al.* 1971, Chs. 4-5, Suppes 1974, Davidson 1990, Weirich 2015, 46). In all such cases, representation theorems are employed to show how sufficiently rich evidence regarding behavioural preference patterns can be used to empirically constrain the range of credence and utility states that an agent might be in. We might call this a *measurement application* of a representation theorem.

Ramsey's general strategy was to assume that CEU is *descriptively* accurate with respect to an agent S's decision-making procedure. Given then that we can empirically ascertain S's preferences, Ramsey proposed to determine her credences and utilities using the CEU representation theorem that he developed. That is, if S is preference-rational with respect to his theorem's conditions C, then according to the Decision-theoretic Interpretation of that theorem and in light of its Standard Uniqueness Condition, S would be a probabilistically coherent expected utility maximiser *only if* she has credences $\mathcal{B}el$ and utilities $\mathcal{D}es$—there is only one probability function $\mathcal{B}el$ which can give rise to her preferences according to CEU, and the only possible $\mathcal{D}es$ functions compatible with her preferences are positive linear transformations of one another. Since we began with the assumption that S does conform to CEU, it follows immediately that we can be confident that S has credences $\mathcal{B}el$ and utilities $\mathcal{D}es$. To the extent that his initial assumption was justified, Ramsey's theorem appears to give us a way to *work backwards* from knowledge of preferences to knowledge of credences and utilities.

Ramsey's measurement system is a prime example of how representation theorems—especially those with the Standard Uniqueness Condition—can help supply us with a solution to the classic *problem of separability*, wherein two distinct quantities usually only have observable consequences when in interaction with one another—thus posing the problem of how to disentangle their respective influences in order to supply a measure for each quantity. In the present situation, this problem is particularly pronounced: according to folk psychology, the main effects of credences—i.e., preferences and intentional action—are only manifest when they interact with utilities, and *vice versa*. As Davidson puts the problem,

> If a person's [utilities] for outcomes were known, then his choices among courses of action would reveal his credence; and if his credence [sic] were known, his choices would disclose the comparative value he puts on the outcomes. But how can both unknowns be determined from simple choices or preferences alone? (1990, 316-7)

For instance, consider the following experiment. An ordinary playing card is placed face-down on a table in front of a subject *S*. No information is given about which card it is. The experimenter gives the subject two choices:

(a) A banana if the card is numbered; an apple otherwise
(b) An apple if the card is numbered; a banana otherwise

Suppose that *S* chooses (a). The problem for the experimenter is to determine why *S* made this choice. Two hypotheses are immediately apparent, each of which presuppose that *S* is maximising her expected utility: either she prefers bananas to apples and is more confident that the card is numbered than that it's not; or, she prefers apples to bananas and is more confident that the card is not numbered. The choice of (a) over (b) does not provide any clear evidence for one hypothesis over the other, and yet the two hypotheses offer contradictory claims about *S*'s credences and utilities. Much of the appeal of many representation theorems with the Standard Uniqueness Condition originates with their apparent capacity to solve this problem—with enough information surrounding the agent's preferences, these theorems suggest that we can narrow down the range of competing hypotheses to what is in effect a *unique* model of the agent's credences and utilities.

So much for the measurement application. Note that while it involves a commitment to the *epistemological* thesis that preferences provide information about credences and utilities, the use of representation theorems in this capacity does not carry any commitment to the *metaphysical* thesis that credences and utilities are characterisable largely in terms of preferences. Historically, however, characterisational representationism has been only a small step on from a Ramseyan measurement application (though perhaps a giant leap for philosophers).

Many historical proponents of characterisational representationism have been sympathetic to some form of operationalism and/or behaviourism with regards the psychological attributes. Ramsey himself seems to have wanted his preference conditions to underlie both a measurement system and a characterisation of credences, asserting that the notion of graded belief "has no precise meaning unless we specify more exactly how it is to be measured" (1931, 167). The main difference between characterisational representationism and the measurement application is that, according to the former, preferences don't just supply good *evidence* about credences and utilities—rather, having (or being disposed to have) appropriate preference patterns is in some important sense a part of *what it is* to have credences and utilities.

3.2 Bridging representation and reality

Let us begin with an outline of a very simple version of characterisational representationism, based on a CEU theorem. Given that an agent is preference-rational with respect to some CEU theorem's conditions C, and given that the theorem has a sufficiently strong uniqueness condition, we can apply similar reasoning as Ramsey's to work backwards from those preferences to a paramorphic model of the agent as an expected utility maximiser with such-and-such credences and utilities. As many authors have noted, however, there is a large gap between the claim that S can be *paramorphically* modelled in a certain way, and the claim that S can be *homomorphically* modelled in that way.[20] Thus, we will need to bridge that gap: a story will need to be told about the connection between representation and reality.

As it turns out, there are many stories that characterisational representationists might tell. Here is a very naïve approach to bridging the gap:

> **Naïve Characterisational Representationism**
> If S can be paramorphically modelled as following some decision rule under a set of credences $\mathcal{B}el$ and utilities $\mathcal{D}es$, then S has credences $\mathcal{B}el$ and utilities $\mathcal{D}es$

If Naïve Characterisational Representationism were true, then every paramorphic model of S would be, *ipso facto*, a homomorphic model. In the event that S satisfies the preference conditions associated with some CEU theorem, the Decision-theoretic Interpretation of that theorem tells us that S can be paramorphically modelled as an expected utility

[20] This has been noted, amongst others, by Maher (1993), Zynda (2000), Christensen (2001), Eriksson and Hájek (2007), Hájek (2008), and Meacham and Weisberg (2011).

maximiser with credences $\mathcal{B}el$ and utilities $\mathcal{D}es$—in which case, Naïve Characterisational Representationism implies that S has credences $\mathcal{B}el$ and utilities $\mathcal{D}es$.

Naïve Characterisational Representationism is deeply flawed, for reasons pointed out by Lyle Zynda (2000). It is easy to show that whenever S can be paramorphically modelled as having credences $\mathcal{B}el$ and utilities $\mathcal{D}es$ combined in a particular way according to a particular modelling scheme Ψ, there will *also* exist another, *radically different* modelling scheme Ψ* that models S as having a *different* set of credences and/or utilities, *combined according to some other decision rule*—indeed, there will usually be *infinitely* many alternative modelling schemes. Simply put, one can make changes in $\mathcal{B}el$ and $\mathcal{D}es$ which are compensated for by changes in the decision rule so as to ultimately produce the same overall pattern of preferences. In light of this, Naïve Characterisational Representationism ends up committing us to assigning to S multiple, radically different and jointly inconsistent sets of credences and utilities.

On one way of looking at it, Zynda's point is one which folk psychologists and philosophers of mind have known for a long time: there are many different possible psychological processes which could underlie a given system of preferences. There are, therefore, many ways to paramorphically represent agents *qua* decision-makers. Note again that this is not in conflict with the theorem having the Standard Uniqueness Condition: as noted in §2.4, uniqueness conditions only specify the range of possible $\mathcal{B}el$-$\mathcal{D}es$ pairs compatible with the agents' preferences *given* that $\mathcal{B}el$ and $\mathcal{D}es$ have certain properties and that they are combined in a certain way.

Taking this into account, another approach to bridging representation and reality might go along the following lines, where Ψ is a modelling scheme established by the Decision-theoretic Interpretation of a CEU theorem:

> **Classical Characterisational Representationism (CCR)**
> If S can be paramorphically modelled *under the modelling scheme* Ψ as an expected utility maximiser with credences $\mathcal{B}el$ and utilities $\mathcal{D}es$, then S has credences $\mathcal{B}el$ and utilities $\mathcal{D}es$

Classical Characterisational Representationism (or CCR) avoids Zynda's worry by relativising to Ψ. The existence of alternative schemes like Ψ* are, on this picture, irrelevant: S's credences and utilities are to be characterised using Ψ, not using some arbitrary modelling scheme Ψ*. Of course, CCR would stand in need of some justification for focusing on Ψ rather than Ψ*—perhaps one could argue that the Ψ scheme is simpler or more natural, being more in line with our pre-theoretic attributions of credences and utilities to ourselves and others.[21] Or, perhaps Ψ and Ψ* should be understood as mere *notational variants*, with the choice to focus on Ψ being a matter of convention (cf. Zynda 2000, Meacham and Weisberg 2011, 657-60). As Davidson (1991, 210ff) argues, there may be

[21] On this idea, see §4.2 below.

nothing more to the possibility of multiple modelling schemes than the common phenomenon of *scale change*, seen for example in the fact that there are infinitely many ways (or scales) by which to represent the weight, length, or temperature of an object, the choice between which is largely a matter of convention.

I refer to this as the 'classical' version of characterisational representationism, though it will not be found anywhere in the literature *quite* as it has been stated here. I am inclined to think that the relativisation to a particular modelling scheme is implicit in most discussions by proponents of characterisational representationism—though the reliance on a CEU theorem in particular is often made very explicit. That any agent with any set of preferences can be paramorphically modelled in innumerable ways is pre-theoretically obvious—the charitable position to take is that historical proponents of characterisational representationism were restricting their attention to very specific ways (modelling schemes) by which to model the preference-rational agent—i.e., as a probabilistically coherent expected utility maximiser. This is made somewhat apparent, for example, by Frank Ramsey (see §7.1) and in Patrick Maher's work (discussed more below).

Note, though, that CCR does rely on the CEU theorem in question having at least the Standard Uniqueness Condition: $\mathcal{B}el$ must be unique, and $\mathcal{D}es$ unique up to positive linear transformations. Without this, the theorem would establish the existence of far too many modelling schemes, involving distinct and incompatible $\mathcal{B}el$ functions, but where *each* such scheme would have us model *S* as an expected utility maximiser. It may be possible to justify a preference for one modelling scheme Ψ under which *S* maximises expected utility over another scheme Ψ^* under which *S* is modelled as following a relatively unintuitive decision rule—but where there are multiple ways to model *S* as an expected utility maximiser, the choice of one way over the other would seem arbitrary at best.

CCR is a conditional claim, leaving open what we might say in the event that *S* does not satisfy the theorem's preference conditions; it only asserts that being modellable in a particular way is *sufficient* for having a particular set of credences and utilities. I suspect that most historical proponents of characterisational representationism would reject the idea that the satisfaction of their favourite CEU theorem's preference conditions is a *necessary* condition for having credences and utilities. In particular, they would reject the following:

> **Extreme Characterisational Representationism**
> *S* has credences $\mathcal{B}el$ and utilities $\mathcal{D}es$ iff *S* can be paramorphically modelled under the modelling scheme Ψ as an expected utility maximiser with credences $\mathcal{B}el$ and utilities $\mathcal{D}es$

For example, Savage is clear that the conditions of his theorem are only meant to characterise a "highly idealised subject" (see his 1954, 5-7). Savage clearly supposes that ordinary, non-ideal folk also have reasonably specific credences and utilities with regards a great many propositions—the instrumental value of his decision-theoretic framework requires as much—so one can charitably assume that Savage didn't believe that his preference conditions were *necessary* for having credences and utilities.

For a complete account of the nature of the graded attitudes, a proponent of CCR would need to give some story for agents who don't satisfy the relevant theorem's conditions (*if* it's supposed that such agents have credences and utilities at all).[22] A first pass suggestion along these lines would be to appeal to what *would* be the case *were* the agent to satisfy the conditions:

> S has credences $\mathcal{Bel}$ and utilities $\mathcal{Des}$ iff, if S were preference-rational, then S would be paramorphically modelled, by Ψ, as an expected utility maximiser with credences $\mathcal{Bel}$ and utilities $\mathcal{Des}$

Pettit (1991) seems to suggest something along these lines, and the idea is critiqued by Meacham and Weisberg (2011, 650-1). A nearby—and I think, more plausible—suggestion would be to characterise an agent's credences and utilities using the representations assigned to the preference-rational agent(s) that they most closely *approximate*:[23]

> S has credences $\mathcal{Bel}$ and utilities $\mathcal{Des}$ iff (i) S approximates at least one preference-rational agent, and (ii) the preference-rational agent(s) that S most closely approximates can be paramorphically modelled, by Ψ, as an expected utility maximiser with credences $\mathcal{Bel}$ and utilities $\mathcal{Des}$

To the extent that the ordinary agent does not even come *close* to being preference-rational—for instance, if the theorem's preference conditions were excessively demanding and unrealistic—the two foregoing suggestions seem implausible. After all, why should the preferences of some hypothetical preference-rational agent S* matter for the determination of S's own credences and utilities, if S* is not at all similar to S? In any case, the key point to recognise for our purposes is that both of these suggestions end up implying CCR: If S *is* preference-rational, then she most closely approximates herself, and the nearest possible world where she is preference-rational is the actual world—so in either case Ψ would model her as having credences $\mathcal{Bel}$ and utilities $\mathcal{Des}$.

[22] Furthermore, an account would be needed for credences and/or utilities towards any propositions P which fall outside of the domains of $\mathcal{Bel}$ and $\mathcal{Des}$, if any such credence/utility states exist. Let us set aside issues relating to this point for now.

[23] In order to develop this latter suggestion, one would of course need a measure of the degree to which one system of preferences approximates another, a statement of when S approximates a preference-rational agent *enough*, and something to say in the event that S most closely approximates more than one preference-rational agent. I am intentionally leaving the notion of *closeness* (or *approximation*) with respect to satisfying preference conditions intuitive and vague—the points I wish to make do not depend on the details of any specific measure.

Patrick Maher—perhaps the most visible proponent of characterisational representationism in recent decades—provides an account of credences and utilities that is designed to apply to all agents, not just those who satisfy the preference conditions of his CEU theorem (see esp. his 1993). Maher adopts an *interpretivist* picture very similar to Lewis' position in his (1974)—discussed in more detail in §4.2—according to which:

> An attribution of [degrees of belief] and utilities is correct just in case it is part of an overall interpretation of a person's preferences that makes sufficiently good sense of them and better sense than any competing interpretation does. (1993, 9)

Maher argues, however, that an interpretation of an agent's preferences which treats her as a probabilistically coherent expected utility maximiser is, in all cases, a *perfect* interpretation:

> [I]f a person's preferences all maximise expected utility relative to some [probability function $\mathcal{Bel}$] and [$\mathcal{Des}$], then it provides a perfect interpretation of the person's preferences to say that [$\mathcal{Bel}$] and [$\mathcal{Des}$] are the person's [credence] and utility functions. (1993, 9)

Maher also implicitly assumes that any such interpretation is *uniquely* perfect (cf. Hájek 2008, 805-6)—and given this, his view ultimately implies CCR. Note, though, that it does not imply Extreme Characterisational Representationism: Maher does *not* assume that it is generally the case that ordinary agents abide by his theorem's preference conditions, nor does he assume that any ordinary agent is best interpreted as a probabilistically coherent expected utility maximiser. It is consistent with his account that no ordinary human being is *ever* preference-rational in the relevant sense.

3.3 Problems with Classical Characterisational Representationism

Naïve, Extreme, and Classical Characterisational Representationism have been the main focus of criticisms against characterisational representationism. In a recent critical review, Meacham and Weisberg (2011) present a number of arguments against five variations on the basic theme of characterisational representationism, *each* of which entail the classical version—and the majority of problems that they raise result from this entailment (and are summarised below). The same can be said more generally: Objections to characterisational representationism often come in the form of arguments against either Naïve, Extreme, or Classical Characterisational Representationism. See, for example, the critical arguments of Hampton (1994), Christensen (2001), Hájek and Eriksson (2007), Hájek (2008), and Easwaran (2014).

In the rest of this section, I will discuss the main concerns which have been raised in the literature (along with some additional concerns of my own). These I have divided into two classes: Those which arise from the very strong connections that these views posit between preferences, credences, and utilities (§3.3.1), and those which arise from the use of CEU theorems in particular (§3.3.2).

3.3.1 The connection with preferences

There are, I think, good reasons to reject CCR and any view which entails it. Perhaps the most common concern is that CCR suggests (and is often held by those with) an anti-realist stance towards the graded propositional attitudes. That is, CCR implies that having certain preferences is *sufficient* for having such-and-such credences and utilities, as though being in those latter kinds of state were nothing over and above being in a particular kind of preference state. In Joyce's words, CCR can "foster a kind of pragmatism that sees belief [or credence] as a second-class propositional attitude that can only be understood in terms of its relationship to desire [or preference]" (1999, 89).[24] The worry is even more apparent with Extreme Characterisational Representationism, or with the suggested approximation-based extension of CCR outlined above—according to either view, to have credences and utilities at all *just is* to have a particular pattern of preferences.

Worse still, where a theorem's $\succcurlyeq$ is understood behaviouristically (as it's usually intended to be), these positions suggest an outdated form of behaviourism—that there is nothing more to having credences and utilities than behaving (or being disposed to behave) in a particular kind of way. Such a position is contrary to our shared, pre-theoretic conception of these things, where our credences and basic utilities for outcomes are understood as each playing a part in the *causal explanation* of our choices. On intuitive grounds, this strongly suggests that credences, utilities, and preferences (whether understood mentalistically or behaviourally) should be kept conceptually and metaphysically separated (cf. Joyce 1999, 21-2).

There is, I think, another problem here, and one which goes beyond a simple knee-jerk reaction to the anti-realism or behaviourism suggested by CCR. The (necessary) preference conditions of a CEU theorem are most plausibly read as *normative* constraints on preferences. *Descriptively*, however, ordinary agents frequently fall foul of basic norms of rationality (whether for systematic or non-systematic reasons), and this creates problems for CCR. Importantly, it's immensely plausible that ordinary agents will sometimes *fail* to have preferences that maximise their expected utility, given their credences and utilities. This means, for one thing, that an agent might have probabilistically coherent credences $\mathcal{Bel}$ and utilities $\mathcal{Des}$ but *not* have expected utility maximising preferences.

[24] Note that how Joyce *officially* defines 'pragmatism' makes it the meta-normative claim that epistemic norms are grounded in practical norms (the "laws of desire"). This kind of view perhaps makes the most sense under an anti-realist or behaviouristic construal of credences in terms of behavioural preferences, but it need not be committed to those construals.

More importantly, it means that an agent might, due to some irrational state of mind, have preferences which could be paramorphically modelled as maximising expected utility relative to some $\mathcal{Bel}$-$\mathcal{Des}$ pair despite not having credences $\mathcal{Bel}$ or utilities $\mathcal{Des}$. In effect, CCR allows for no wriggle room between preferences on the one hand and credences and utilities on the other, in the event that those preferences satisfy the relevant theorem's conditions. It implies that it's *impossible* for anyone to be preference-rational *by accident*—that, whenever someone's preferences conform to the conditions, it *must* be because they were acting rationally given their credences and utilities. CCR implies that irrational agents *cannot* satisfy a CEU theorem's preference conditions, and this seems utterly unmotivated.

There are some obvious changes that one could make to CCR to loosen the connection it posits between preferences on the one hand, and credences and utilities on the other. For reasons that I will return to shortly, I doubt that these will be quite enough, but it's worth highlighting them briefly first. To begin with, we might contrast CCR with the following account, inspired by Lewis (1980a):

> S has credences $\mathcal{Bel}$ and utilities $\mathcal{Des}$ iff S is in some psychophysical state R^*, where R^* would *tend to cause* a typical subject S' to be preference-rational such that she would be modelled, by Ψ, as an expected utility maximiser with credences $\mathcal{Bel}$ and utilities $\mathcal{Des}$

This kind of view would require that credence states are identifiable independently of their functional role—i.e., as a particular *neurobiological kind*. Like Lewis, one might cash out the *tends to cause* relation by reference to the causal role that R^* would play in a typical member of some pre-specified population. There are, however, other ways to flesh out the relation, which we need not consider here; the important point for our purposes is that it's not CCR: the fact that an underlying psychophysical state R^* *tends to cause* preference-rationality does not mean that whenever the agent is preference-rational, they are therefore in R^*. Perhaps CCR holds much of the time, or holds for a perfectly typical subject, but it need not hold in general.

The second way in which CCR might be avoided would be to ignore *actual* preferences and instead characterise an agent's credences and utilities in terms of what preferences she *would* have in some idealised state. We have noted that the ordinary subject will often make mistakes, in one way or another failing to have the pragmatically optimal preferences given her credences and utilities. However, perhaps under some idealised state of considered reflection, every agent will conform to decision-theoretic norms:

> S has credences $\mathcal{Bel}$ and utilities $\mathcal{Des}$ iff where S in ideal conditions (e.g., she is functioning properly in a normal environment, free from interfering influences such as intoxication, time pressures, and so on), then S would be paramorphically modelled, by Ψ, as an expected utility maximiser with credences $\mathcal{Bel}$ and utilities $\mathcal{Des}$

I suspect that something like this is probably true (see §8.5), but as an account of the nature of credences it still seems to be missing something. While it may be plausible that *utilities* straightforwardly reduce to particular patterns of preferences (especially where 'preference' is given a mentalistic construal), our *credences* seem to be a wholly distinct and independently existing kind of mental state—and the above two suggestions do not yet capture everything which is important about them.

Importantly, the credences that we have towards specific propositions seems to depend strongly on the evidence that we have accumulated regarding to those propositions. However, there is no accommodation for this connection between credences and past evidence in CCR (or any of the proposed refinements). The worry here is expressed nicely in the following passage by David Christensen (see also Weirich 2004, 20, for similar remarks):

> True, degrees of belief are intimately connected with preferences and choice behaviour. But they are also massively and intimately connected with all sorts of other aspects of our psychology (and perhaps even physiology). This being so, the move of settling on just one of those connections—even an important one—as definitional comes to look highly suspicious. (2001, 362)

Building off of Christensen's discussion, Meacham and Weisberg make the same complaint:

> Given that beliefs have connections to so many mental states besides preference—emotions, perception, memory, and so on—it's implausible that just one of these connections is paramount. With all the pushes and pulls that beliefs and desires are entangled in, we should not expect there to be a rigid and straightforward connection between degrees of belief, utility, and preference. (2011, 646)

Indeed, CCR could have us assign credences $\mathcal{B}el$ to an agent on the basis of her preferences even when $\mathcal{B}el$ is *entirely* at odds with what we would expect her credences to be like given her life history of evidence. And *this* result seems unacceptable.

3.3.2 The use of CEU theorems

Another frequent cause for concern arises from the use of CEU theorems in particular. The focus on CEU theorems is, I suspect, due largely to the attention philosophers have given to characterising the credences of ideally rational agents. The use of CEU theorems has been the grounds of two basic criticisms, which I will discuss in turn.

The first criticism is that ordinary agents do not satisfy the preference conditions associated with standard CEU theorems. This complaint plays a prominent role in the critical discussions found in (Hampton 1994), (Meacham and Weisberg 2011), and (Dogramaci forthcoming). Much of the relevant empirical work is summarised in (Tversky 1975), (Camerer 1995), (Schmidt 2002), and (Johanna, Jeleva *et al.* 2012). The most widely cited evidence here originates with Allais (1953) and Ellsberg (1961). Kahneman and Tversky (1979) outline experimental results which (they argue) imply that ordinary decision makers in the kinds of decision situations that Allais outlined don't always adhere to Savage's *sure-thing principle*, which is a common independence condition found in many CEU theorems (see §5.1.2). The adequacy of other independence conditions also comes under attack from (Birnbaum and Chavez 1997) and (Birnbaum and Beeghley 1997). Some authors have also purported to show through so-called preference reversal experiments that ordinary agents' preferences are intransitive (Lichtenstein and Slovic 1971, 1973, Fishburn and LaValle 1988).[25]

It is perhaps not so worrying if most agents don't satisfy the CEU conditions *exactly*, so long as they come *close* to satisfying those conditions (in which case we might appeal to the preference-rational systems they most closely approximate). One reason to think that ordinary agents' preference don't vary greatly—or at least, greatly *and* systematically—from CEU-consistent preferences is that many predictive models in economics and the social sciences essentially treat the average decision-maker as having the kinds of preferences associated with expected utility maximisers, *or thereabouts*.[26]

Even descriptive decision-theoretic models that are explicitly designed to accommodate the empirical evidence for our deviations from CEU bear a close resemblance to that theory: with few exceptions, they involve a $\mathcal{B}el$ function (which is at least a capacity if not a probability function) and a utility function combined in something like expectational form, with the basic decision-making principle being that an agent will pick the option which has the highest $\mathcal{B}el$-weighted average utility. This is essentially the case, for example, of Kahneman and Tversky's (1992) *cumulative prospect theory*, which is widely considered to be the most empirically accurate decision model so far developed. (See Appendix B for more details.)

[25] The vast majority of the empirical work has focused on whether ordinary agents satisfy the *necessary* conditions associated with CEU theorems; whether they always satisfy the non-necessary, structural conditions is generally taken to be relatively unimportant. The main reason for this attitude will be discussed in §5.2.4.

[26] The general point here goes back at least to Fodor (1987), who argued that folk psychology (which is in many respects very close to orthodox expected utility theory) is presupposed so widely within our best explanations of behaviour that it is likely to be at least broadly correct.

There may be some bells and whistles added, or some relatively minor deviations from strict requirements of rationality here and there, but—as a rule—NCU models of decision-making are generalisations of the traditional CEU models. Consequently, the preference conditions which underlie NCU representation theorems tend to be weaker versions of the conditions underlying CEU theorems.[27] It would be a mistake to infer from the apparently vast amount of evidence that we don't satisfy the preference conditions for a CEU theorem that we are therefore *far* from satisfying those conditions. If anything, the evidence that we have suggests exactly the opposite conclusion.

The second concern that commonly arises from the specific appeal to CEU theorems concerns the *empirical plausibility* of the decision-making models that CEU theorems generate. There are two sub-issues to distinguish here. The first concerns whether ordinary agents can be plausibly understood as expected utility maximisers, especially given the range of alternative psychological models of our decision-making processes. As I've just noted, these other models do tend to be very *similar* in their broad structure to classical expected utility theory—but the point is nevertheless sound: ordinary agents are probably *not* expected utility maximisers across the board.[28]

The second (and closely related) sub-issue results from the fact that CEU theorems are fundamentally limited in their capacity to represent credence states, requiring as they do that $\mathcal{B}el$ is a probability function. The complaint, of course, is that if ordinary agents are not probabilistically coherent then *no* probability function can faithfully model her total credence state—and there are many ways that one could fail to be probabilistically coherent. A version of this complaint has been raised in most critical discussions of characterisational representationism.

There is an important background assumption being made here, which is that the $\mathcal{B}el$ of any ordinary CEU theorem must be understood as being defined on some algebra of sets $\boldsymbol{\mathcal{P}}$ defined on a space of *possible* worlds, $\boldsymbol{\mathcal{W}}$ (or in the case of Savage-like theorems, a space of *possible* states of affairs, $\boldsymbol{\mathcal{S}}$). For any such probability function, the following conditions will hold:

(i) *Logical omniscience*: $\mathcal{B}el(\top) = 1$ for any necessary proposition $\top \in \boldsymbol{\mathcal{P}}$, and $\mathcal{B}el(\bot) = 0$ for any impossible proposition $\bot \in \boldsymbol{\mathcal{P}}$

(ii) *Additivity*: $\mathcal{B}el(P \,\&\, Q) = \mathcal{B}el(P) + \mathcal{B}el(Q)$, for any logically incompatible pair of propositions $P, Q \in \boldsymbol{\mathcal{P}}$

[27] More specifically, many NCU theorems essentially result from various ways of weakening the preference conditions found in Savage's CEU theorem.

[28] It is hard, however, to find *any* proponent of characterisational representationism whose view commits them to asserting that all agents are at all times expected utility maximisers—that commitment doesn't follow from CCR, but requires something more like Extreme Characterisational Representationism.

(iii) *Monotonicity*: If P, $Q \in \mathcal{P}$, then $\mathcal{B}el(P) \geq \mathcal{B}el(Q)$ if $Q \vdash P$ (corollary: if P and Q are logically equivalent, then $\mathcal{B}el(P) = \mathcal{B}el(Q)$)[29]

Individually, each of (i) to (iii) seems an implausible condition to impose upon a model of an ordinary agent's credences; they only seem reasonable for deductively infallible, hyper-rational beings, who recognise all the logical implications of every proposition they contemplate. (i) is clearly too strong: there are many logical or mathematical truths of which I am not certain, and many logical or mathematical falsehoods to which I give some positive credence. There is, furthermore, a wealth of empirical evidence against the descriptive plausibility of both (ii) and (iii), which I will not repeat here—though see especially (Tversky and Kahneman 1974).

Philosophers routinely assume that failures of (i) to (iii) are *typical* of ordinary agents.[30] (Otherwise, there would be little point in arguing so much over whether agents *ought* to be probabilistically coherent!) If this is true, and ordinary agents are generally and sometimes quite drastically probabilistically *incoherent*, then we have a clear problem for any version of characterisational representationism based solely on a CEU theorem, *where* $\mathcal{B}el$ is defined on an algebra constructed from a space of *possibilities*.

In §4.3, I will show that a probability function defined on a space of possible *and* impossible worlds (or states) need not satisfy any of the conditions (i) to (iii). The important question, though, is whether the Decision-theoretic Interpretation of any CEU theorem is compatible with this understanding of the domain of its probability function. There seems to be no good reason for supposing that Jeffrey's set $\mathcal{W}$ must be composed of possible worlds only. On the other hand, it's less clear whether letting Savage's set $\mathcal{S}$ include impossible states sits well with the rest of his framework. Further discussion of these issues, however, will have to await a more detailed look at the relevant theorems.

3.4 Desiderata

We have seen three broad kinds of complaints that have been raised against CCR and the positions which imply it. First of all—and, I think, most importantly—it's very plausible that credences and utilities ought to be kept metaphysically and conceptually distinct from preferences. Having a credence of x in P, in particular, is *not* just a kind of preference

[29] Given (i), monotonicity is of course implied by additivity; however, it is useful to distinguish the two properties here—particularly because Choquet capacities satisfy monotonicity without always satisfying additivity.

[30] There is some doubt on this front—see especially (Lewis 1982, 1986, 34-6) and (Stalnaker 1984). I have neglected to discuss their proposed solutions to 'the problem of logical omniscience' here as accepting that ordinary agents' credences satisfy (i) to (iii) comes at a high intuitive cost, and (moreover) because I do not think characterisational representationism is committed to the idea that ordinary agents must be probabilistically coherent.

state. There needs to be some wriggle room between the two kinds of states; credences appear to *give rise to* preferences, but not with invariable certainty, and not always through expected utility maximisation. Furthermore, credences seem to play other roles besides their role in the production of preferences—for instance, they change in response to evidence—and this needs to be accounted for in any adequate characterisation of what it is to have credences. One of the central roles of Chapter 4 is to show that characterisational representationism can accommodate this lesson.

The second kind of complaint arises from the details of the theorems which underlie CCR. These decompose into two basic issues: first, whether ordinary agents satisfy (or come sufficiently close to satisfying) the relevant theorems' preference conditions; and second, whether the ensuing models of their credences, utilities, and decision-making process are plausible.

Supposing that we were to remove CEU theorems from consideration, the obvious alternative for characterisational representationism would be to appeal to some NCU theorem. NCU theorems are, for the most part, explicitly designed to capture the preference patterns of ordinary agents, and many of them allow for *non*-probabilistic credence functions. Typically, NCU theorems achieve this with weaker preference conditions than those found in CEU theorems—that is, preferences which satisfy the CEU conditions will also satisfy the NCU conditions, but not *vice versa*. By setting weaker and more realistic preference conditions and allowing for a broader range of representations, NCU theorems should look like a very attractive place to search for a firmer basis for characterisational representationism.

Of course, not *any* NCU theorem will do—we need a theorem with *the right properties*. To close this chapter, then, I want to say in more general terms what kinds of features a representation theorem *should* have, *if* it is to underlie a more plausible version of characterisational representationism. I will begin with a very schematic discussion of some basic conditions on any characterisation of credences and utilities.

Any minimally realist account of what it takes to have such-and-such credences and utilities can be put very schematically as follows:

> S has credences $\mathcal{B}el$ and utilities $\mathcal{D}es$ iff S satisfies conditions N

Now, first of all, for any plausible account it ought to be the case that:

> **Satisfiability**
> Ordinary agents generally satisfy conditions N

If Satisfiability is not satisfied, then the account is unable to explain how it is that ordinary agents have the credences and utilities that they do. Secondly, the account ought to be plausible:

> **Plausibility**
> The $\mathcal{B}el$ and $\mathcal{D}es$ assigned to S under conditions N are plausible models of S's credences and utilities, in the sense that they broadly coincide with our intuitions/empirical data regarding what credences and utilities an ordinary agent would have in those conditions

If Plausibility is not satisfied, then we have good reason to think that the account is not picking out those intentional states which we understand to be credence and utility states, nor anything in the vicinity. Note that perfect fit with our intuitive judgements is not required to satisfy this second condition: there is always some room to move away from what is intuitive, if such manoeuvres are well motivated.

Thirdly, the account should not be circular:

> **Non-Circularity**
> N ought to specifiable without reference to S's credences or utilities

If Non-Circularity is not satisfied, then the characterisation is obviously at least somewhat circular, requiring prior knowledge of S's credences and/or utilities before being able to specify what her credences and utilities are.

Furthermore, if the goal is to generate a fully naturalistic account, then the following should be satisfied:

> **Naturalisability**
> N ought to be specified by reference only to natural or readily naturalisable properties

Paradigm instances of concepts which are *not* readily naturalisable include the semantic, intentional, and other mental concepts—hence the focus on naturalising intentionality, rather than (say) the property of *being a teacup*.

The advocate of characterisational representationism seeks to make essential use of a representation theorem in her characterisation of credences and utilities. Specifically, she thinks that the *modelling schemes* generated by such a theorem can tell us a great deal about what an ordinary agent's credences and utilities *are*, and how she forms her decisions, given enough information about the agent's preferences (in either the mentalistic or behavioural sense). As I will argue in Chapter 4, there need not be any simple and straightforward relationship between how an agent can be modelled according to the theorem and what her mental states actually are—CCR is not the only game in town—but there must nevertheless be a close relationship between the modelling scheme and the mental facts of the matter, *if* the view is to count as an instance of characterisational representationism.

Our three conditions, Satisfiability, Plausibility, and Non-Circularity, therefore, can be used to generate a number of basic desiderata that we might want a representation theorem to satisfy, if it is to provide a plausible foundation for the characterisational representationist's project. Naturalisability, likewise, generates a further desideratum for naturalisers. These desiderata are summarised in §3.4.5 below.

3.4.1 The theorem's condition's satisfiability

Given Satisfiability, it's clear that a representation theorem will only be useful for the purposes of characterisational representationism to the extent that its preference conditions are satisfied—or at least approximately satisfied—by ordinary agents, at least under appropriately specified conditions.

An absolutely minimal requirement, then, is that the theorem's preference conditions are (approximately) *satisfiable*. (In what follows, I will set aside the 'approximately' qualifier for ease of reading.) As we will see in §5.2.1–2 and to a lesser extent §7.2.2, we should not take it for granted that the conditions placed on $\langle \mathcal{BOP}, \succcurlyeq \rangle$ can be satisfied by *any* agent's preference system. As noted in §2.4, the fact that $\langle \mathcal{BOP}, \succcurlyeq \rangle$ is *intended* to represent a possible preference system is not sufficient reason to assume that it does so. Formally characterised, $\succcurlyeq$ and/or $\mathcal{BOP}$ may have properties which would make no sense under the intended interpretation. In particular, $\succcurlyeq$ may end up being defined on a collection of entities which bear no resemblance to what might reasonably be considered a set of basic objects of preference, under any conception of 'preference'.

Supposing, then, that the representation theorem's conditions are satisfiable, we can consider whether the conditions *are* satisfied. Characterisational representationism, based on a given representation theorem T, is more plausible to the extent that T's preference conditions actually *are* satisfied by the average person on the street, at least under appropriately specified conditions. If no ordinary agent (in the right conditions) ever satisfies the T's conditions for representation, even approximately, then the theorem would seem to have nothing interesting to tell us about the credences and utilities of ordinary agents.

To be sure, if some kind of idealised being were to satisfy the conditions, then the theorem may have something to say about *their* credences and utilities—but characterisational representationism is a thesis about the credences and utilities *in general*, not just the mental states of an idealised subject who (by hypothesis) does not even come close to having an ordinary preference system. I will discuss this point further below.

3.4.2 The plausibility of $\mathcal{B}el$, $\mathcal{D}es$, and the decision rule

To the extent that a theorem's modelling scheme is to provide information towards agents' actual total credence and utility states, its $\mathcal{B}el$ and $\mathcal{D}es$ ought to be adequate *qua* models of those states.[31] Furthermore, the overall model of the agent *qua* decision-maker should be plausible, in light of what we know about ordinary agents and how they make decisions.

For instance, suppose that S satisfies theorem T's preference conditions, such that T's implies she can be paramorphically modelled as having credences $\mathcal{B}el$ and utilities $\mathcal{D}es$ combined according to some rule $\mathcal{R}$. Such information only seems valuable for the characterisational representationist if $\mathcal{B}el$ and $\mathcal{D}es$ are not wildly at variance with what we would expect S's credences and utilities to be, and $\mathcal{R}$ is not wildly at variance with how we would expect S to form her preferences, under the relevant circumstances. (Even if the $\mathcal{B}el$ and $\mathcal{D}es$ are plausible, if T represents S as an expected utility *minimiser* then something has obviously gone very wrong!) Characterisational representationism is more plausible to the extent that the theorem upon which it is based supplies *plausible* models of our mental states and decision-making procedures; where this doesn't hold, it would seem unreasonable to connect the theorem's $\mathcal{B}el$ and $\mathcal{D}es$ functions in any close way to agents' actual attitudes. The further the theorem's models are away from reality, the less plausible it is that those models might play any important role in characterising the reality.

Let us see if we can specify some more specific requirements for the plausibility of $\mathcal{B}el$ and $\mathcal{D}es$. To begin with, I take it as a conceptual truth that (just as beliefs are relations between an agent and a proposition) credences are relations between agents, propositions, and levels of confidence. So, at minimum, $\mathcal{B}el$ should connect *propositions* (or entities which closely correspond to propositions) to a measure of confidence. It is plausible that $\mathcal{D}es$ ought to take a similar structure—i.e., it ought to connect propositions to a measure of desirability—and throughout this work I have been treating $\mathcal{D}es$ as a mental state with *propositional* content. This presupposition does not figure very heavily in the discussion that follows.

I do assume, however, that the set of potential objects of credence and the set of potential objects of utility are not wholly *disjoint*. That is, it seems reasonable to suppose that any proposition towards which an agent has credences is also a proposition towards which she *could* have utilities, and *vice versa*. There may be a small number of exceptions to this rule, but even granting such exceptions it seems reasonable to require that the domains of our $\mathcal{B}el$ and $\mathcal{D}es$ functions *could* feature a substantial degree of overlap.

[31] Note, again, that these are merely desiderata—just as Plausibility is malleable in the light of other theoretical considerations such as simplicity or theoretical fruitfulness, so too should we allow for some wriggle room between pre-theoretic intuitions and our formal models of credences and utilities.

We can remain non-committal for now on the nature of propositions; for instance, on whether they are sets of (metaphysically/epistemically/conceptually) possible worlds (as in Lewis 1986, Stalnaker 1984, 2008), sets of centred worlds or properties (Lewis 1979, Jackson 2010), structured *n*-tuples of objects and properties (Soames 1987), or otherwise. Whatever the exact nature of propositions, though, it's reasonable to suppose that they ought to be fine-grained enough to capture the hyperintensionality of our credences and utilities. It is plausible, for instance, that one might have a particular degree of confidence in the claim that *water is wet*, without having the same confidence in the claim that *H_2O is wet*. Reasons in favour of this claim are discussed in some detail in (Chalmers 2011) and (Jackson 2009). Similar intuitions suggest that one might have different degrees of confidence towards logically and mathematically equivalent claims; for instance, I am far more certain that $1 + 1 = 2$ than I am in the truth (or falsity) of Goldbach's conjecture. Thus, it is reasonable to expect that $\mathcal{B}el$ and $\mathcal{D}es$ ought to be able to distinguish between metaphysically equivalent—and perhaps even logically and mathematically equivalent—objects of thought.

However we understand propositions, $\mathcal{B}el$ and $\mathcal{D}es$ should also be capable of assigning values to *all* and *only* the propositions that we take ourselves to potentially have credences and utilities towards. To be sure, according to the Decision-theoretic Interpretation, $\mathcal{B}el$ and $\mathcal{D}es$ may only be *partial* models of an agent's credences and utilities respectively—but if characterisational representationism is going to define what it is to have credences and utilities *in general* by means of a given theorem's representation scheme then $\mathcal{B}el$ and $\mathcal{D}es$ had better not leave out too much. (For further arguments on this, see §5.3.1 and §6.1.2.)

It is an interesting (and to my knowledge unsettled) question whether there are any propositions towards which we *cannot* have credences or utilities. I will here briefly consider one suggested restriction; other potential restrictions on $\mathcal{B}el$ and $\mathcal{D}es$ will be discussed in later chapters when they become relevant. In particular, the suggestion I want to consider is that some propositions may be *too complex* to be contemplated (where the *complexity* of a proposition seems to correspond roughly to the complexity of a minimal sentential expression of the proposition in a natural language). If this is so, then we may simply *lack* credences and utilities towards such propositions. However, to the extent that this limitation exists, it seems to only apply to non-ideal agents with limited cognitive capacities; it does not seem to apply to more idealised agents, and presumably characterisational representationism should account even for the ideal case. This suggests that $\mathcal{B}el$ and $\mathcal{D}es$ ought to be *flexible*, in the sense that they should be capable of representing credences and utilities towards highly complex propositions, but also capable of representing an absence of credences and/or utilities towards such things.

Finally, it's worth recalling one of the lessons of §3.3: Where $\mathcal{B}el$ can only take the form of a probability function—or more specifically, a probability function defined on an algebra constructed from a set of *possibilities*—it seems unlikely it could adequately represent the credences that ordinary agents have towards many propositions (even if those agents were to satisfy the theorem's conditions). Similar complaints arise even if $\mathcal{B}el$ is only a Choquet capacity. Such functions are only adequate for logically omniscient agents with monotonic (if not additive) and infinitely precise credences. Furthermore, capacities and probability functions satisfy strong closure conditions; *viz*., they are closed under complementation and under (at least finite) unions.

There is, then, another sense in which $\mathcal{B}el$ ought to be flexible: It should not be limited to probability functions, capacities, or any other kinds of function with excessively restrictive conditions that severely limit their applicability *qua* models of ordinary agents' credences.

3.4.3 The uniqueness condition

A related factor to consider is the strength of the theorem's uniqueness condition. It is pre-theoretically implausible that individual agents' credences and utilities (at a time) can be adequately represented by a wide range of highly divergent (and potentially contradictory) pairs of $\mathcal{B}el$ and $\mathcal{D}es$ functions.[32] To the extent that our credence and utility states are unique, any plausible model of those states ought to be unique.

But this truism does not translate into a requirement that the theorem upon which we base characterisational representationism must have strong uniqueness results. As we will see in Chapter 4, characterisational representationists can appeal to information that goes beyond agents' (actual or counterfactual) preferences, which might be used to narrow down the range of potential interpretations whenever a representation theorem does not deliver strongly unique results. The requirement that we represent an agent as having relatively unique credences and utilities does *not* translate into a requirement that the theorem upon which we base constitutive representationism must have strong uniqueness results. Nevertheless, the extent to which $\mathcal{B}el$ and $\mathcal{D}es$ are unique is an important factor in the evaluation of a representation theorem *qua* basis for characterisational representationism, as the strength of the theorem's uniqueness result impacts upon what kinds of connections can be drawn between the theorem's modelling scheme and the mental facts of the matter.

The large majority of theorems to be considered in the remainder of this work have the Standard Uniqueness Condition; for this reason, the *uniqueness* desideratum does not play

[32] Those who accept that agents can at one time have multiple, fragmented systems of belief might deny this point (cf. Lewis 1982). However, the kind of non-uniqueness that these theorists claim to exist is conceptually quite distinct from the kind of non-uniqueness we are discussing here: these theorists are usually motivated to appeal to fragmented belief states when a single coherent belief state cannot explain an agent's apparently irrational behaviour and preferences, whereas we are now looking at a situation where multiple belief states can each individually be used to explain the agent's behaviour/preferences equally well.

a large role in most of the critical discussions that follow. The major exception is Jeffrey's (1990) theorem where the $\mathcal{Bel}$-$\mathcal{Des}$ pair is only unique up to a fractional linear transformation. I discuss what a characterisational representationist might do with this weaker uniqueness condition in §6.2.2.

3.4.4 The interpretation of the theorem's primitives

The point of characterisational representationism is to define what it is to have such-and-such credences and utilities largely in terms of preferences, by appeal to a representation theorem. If any such project is to be successful, then the basic notions involved in the interpretation of those theorems cannot themselves be understood in terms of credences or utilities. Generally speaking, if the goal is to define X in terms of Y, then it had better not be the case that Y is to be understood, in turn, in terms of X. Thus, from the Non-Circularity condition, we see that if characterisational representationism is to be founded upon some theorem or other, it's a minimal requirement upon that theorem that it can be interpreted without reference to agents' credences and utilities.[33]

There are at least two basic formal elements to any representation theorem: a preference relation $\succcurlyeq$, and a set $\boldsymbol{\mathcal{BOP}}$ of objects of preference. Often, $\boldsymbol{\mathcal{BOP}}$ is itself constructed from a number of further sets. If a given representation theorem is to satisfy the present Non-Circularity condition, then neither $\succcurlyeq$, nor $\boldsymbol{\mathcal{BOP}}$, nor any other primitive elements involved in the statement of the theorem should be given an interpretation which requires reference to credence or utility states. For instance, it would obviously not do for characterisational representationism to define $\succcurlyeq$ as follows:

> $x \succcurlyeq y$ (relative to an agent S) iff S has a higher utility for x than for y

Likewise, suppose that $\boldsymbol{\mathcal{BOP}}$ is supposed to represent a collection of gambles conditional on a proposition P, where it's required that the agent has a particular credence value n for P (e.g., $n = 0.5$). Unless we already know what it is to have credence n in P, preferences over such bets will not be very useful in the characterisation of what it is to be in such-and-such a credence state more generally.

Furthermore, from the Naturalisability condition we know that if we are to provide a naturalistic characterisation of what it is to be in certain credence and utility states, and if

[33] To be clear, some philosophers are happy to countenance non-reductive definitions of important concepts, wherein the definiendum forms part of the definiens. I am assuming, however, that the goal of characterisational representationism (and preference functionalism more generally) is reductive. Recall that much of the appeal that characterisational representationism holds is due to its promise to solve the old philosophical problem that arises from the interdefinability of credence and utility (or belief and desire).

a representation theorem is to play a central role in that characterisation, then the basic notions of the theorem should be naturalistic—or at least readily naturalisable.

3.4.5 Summary

Let us summarise. The following desiderata are important for characterisational representationists generally (whether naturalistic or non-naturalistic); subsidiary desiderata are also listed:

(**1**) The theorem's preference conditions should be *satisfied* (or approximately satisfied) by the majority of ordinary human agents (at least under appropriately specified circumstances).

(**1a**) The theorem's preference conditions must be *satisfiable*.

(**2**) Assuming that *S* is an ordinary agent and satisfies *T*'s preference conditions, *T* should provide a plausible (if slightly idealised) homomorphic model of *S*'s credences, utilities, and preference-forming procedure.

(**2a**) $\mathcal{B}el$ and $\mathcal{D}es$ ought to be capable of assigning values to (more or less) the same propositions, rather than having distinct, non-overlapping domains.

(**2b**) $\mathcal{B}el$ and $\mathcal{D}es$ ought to be capable of modelling hyperintensional credences and utilities—they ought to be capable of distinguishing and assigning distinct values to metaphysically—and perhaps even logically and mathematically—equivalent objects of thought.

(**2c**) $\mathcal{B}el$ and $\mathcal{D}es$ ought to assign values to *all* and *only* the objects of thought towards which the relevant agent has credences and utilities, respectively.

(**2d**) $\mathcal{B}el$ ought to be capable of modelling the total credence states of non-ideal reasoners with potentially indeterminate or imprecise credences; it should not be restricted to models of agents who are probabilistically coherent, logically omniscient, deductively infallible, and so on.

(**2e**) The manner by which $\mathcal{B}el$ and $\mathcal{D}es$ combine to determine preferences should be plausible, under the relevant circumstances.

(**3**) The theorem should have a reasonably strong uniqueness condition.

(**4**) It should be possible to understand and specify the basic notions involved in the interpretation of the theorem independently of any prior knowledge regarding the relevant agents' credences and/or utilities.

Furthermore, if a naturalistic variety of characterisational representationism is the goal, then a further desideratum is:

(**5**) The basic notions of the theorem should be naturalistic/readily naturalisable.

In Chapters 5 through to 7, I will evaluate a range of theorems in light of these desiderata. I will argue that none of them satisfy each of **(1)** to **(4)**; furthermore, I will argue that none satisfy **(5)**.

(1a), **(3)**, and **(4)** seem non-negotiable. However, readers might note the emphasis on *ordinary* agents in **(1)** and **(2)**, and may want to weaken the relevant criteria if their only goal is to characterise credences and utilities for *ideally rational agents*. One might take this as part of a two-step strategy for characterising credences and utilities *in general*: first give an account for the ideal case, and then 'de-idealise' so that it applies to ordinary agents. Taking this line may suggest replacing **(1)** and **(2)** with:

(1′) The theorem's preference conditions should be *satisfied* (or approximately satisfied) by *ideally rational agents* in idealised conditions.

(2′) Assuming that *S* is *ideally rational* and satisfies *T*'s preference conditions, *T* should provide a plausible homomorphic model of *S*'s credences, utilities, and preference-forming procedure.

We can plausibly assume that ideally rational agents are probabilistically coherent, hence adopting **(2′)** might suggest relaxing **(2b)** and **(2d)** in particular. Furthermore, it is plausible that ideally rational agents apply a different decision rule than ordinary agents (or the same rule, but better and more consistently), so **(2e)** would need to be interpreted accordingly.

Something like this two-step strategy for understanding empirical phenomena is applied throughout the sciences, and I strongly suspect that it will be required for present project as well. We should, for instance, certainly focus our attention on properly functioning, species-typical human beings in normal circumstances with slightly idealised cognitive processes free from various, well-known confounding factors (e.g., injury, intoxication, etc.). In *that* sense of 'idealisation', we should indeed attempt to characterise credences and utilities for the ideal case and then see what can be done about de-idealisation. The two-step strategy works best, however, when (i) the relevant idealisations don't leave us vastly removed from the actual, target phenomenon, and (ii) it is reasonably clear how to 'de-idealise'.

What we are after is a characterisation of credences and utilities in general. It's hardly likely, however, that the metaphysics of credences and utilities is *disjunctive*, in the sense of being one way for ideally rational agents and a wholly different way for ordinary agents. So, we should expect any plausible approach to credences and utilities for ideally rational agents to be a special case of a more general account for agents of all kinds. Thus, *if* we are going to develop an account of credences and utilities for ideally rational agents, it should be readily *generalisable*—that is, it should be reasonably clear how to extend (or 'de-idealise') the account so as to apply also to ordinary agents.

What is unclear is whether this ‘de-idealisability’ condition will be met if all we have is a theorem which *merely* satisfies **(1′)** and **(2′)**. Moreover, *showing* that it can be met will essentially involve showing that there is a theorem in the vicinity which satisfies **(1)** and **(2)**. Characterisational representationism won’t be fully vindicated unless progress can be made towards a theorem which satisfies the original desiderata, relevant to the ordinary person on the street. A theorem that applies only to angels is not enough.

In Chapter 6, I will suggest that Jeffrey’s representation theorem may satisfy **(1′)** and **(2′)**, though it does this at the cost of strong uniqueness results. However, the idealisations needed are extreme: Jeffrey’s theorem *only* applies to highly idealised subjects, his representation result is only plausible for the ideally rational agent, and it is not clear whether and how his conditions can be weakened to account for the ordinary subject. In Chapter 8, however, I will suggest an improvement—a theorem which is many respects similar to Jeffrey’s but comes much closer to satisfying **(1)** and **(2)** (as well as **(3)** and **(4)**), and which has the Standard Uniqueness Condition.

CHAPTER FOUR

Interpretivism and Functional Role Semantics

In this chapter, I will argue that with *the right kind* of representation theorem—one which satisfies the desiderata of §3.4.5—there are at least two ways of developing characterisational representationism which avoid the central worries that arise for Naïve, Extreme, and Classical Characterisational Representationism. Each of these ways is closely analogous to an important contemporary account developed for beliefs and desires; to the extent that the latter are taken seriously by philosophers as viable options worthy of development, so too should their counterparts with respect to credences and utilities. My purpose, in other words, is to establish a clear case for pursuing characterisational representationism, and for developing representation theorems with characterisational purposes in mind.

A large part of this chapter will be spent on outlining a variety of positions regarding the nature of propositional attitudes and the manner in which they come to have intentional content. Along the way, I will mark out those areas where representation theorems are likely to prove particularly helpful. I am inclined to favour those positions and will offer some brief arguments against the alternatives, but a criticism of other views is not my focus here. The goal of this discussion is neither completeness nor depth; rather, it is to mark out some of the major positions that a theorist might adopt when it comes to characterising credences and utilities, by way of analogy to some of the major positions which exist in relation to beliefs and desires.

The motivation for this stems ultimately from the fact that the philosophical options for characterising credences and utilities—whether with or without the use of representation theorems—have been left largely unexplored. As a result, there has never been a close investigation into how representation theorems might be applied towards developing an account of the graded attitudes. With the exception of Maher (1993), those friendly to characterisational representationism rarely offer more than a few vague conjectures about how they expect their favoured representation theorem might be of help (usually something along the lines of CCR). Critics of characterisational representationism have done more work in spelling out the options than its proponents have (see especially Meacham and Weisberg 2011, 644-54).

This state of affairs is unfortunate. Perhaps because the relevant terrain of options remains mostly unexplored, characterisational representationism is often quickly dismissed, being labelled a form of *behaviourism* or *anti-realism*—i.e., the kind of views wherein credences and utilities are nothing more than theoretical constructs designed to systematically represent an agent's behaviour. And understandably so: where $\succcurlyeq$ is understood as a kind of behavioural preference, CCR and any view which implies it does strongly suggest a behaviourist and/or anti-realist viewpoint (§3.3.1). Such positions are then placed in contrast with a more wholesome, non-behaviourist and fully Realist (with a capital 'R') perspective, whereby credences and utilities are understood to be genuine, psychologically real states of the agent (in a sense to be specified shortly) with rather more malleable and contingent causal links to choice behaviour (e.g., Weirich 2004, 8, 19-20). Not much more is said about these alternative positions—just that, however it may ultimately be fleshed out, what results will *not* be characterisational representationism (or worse: a betting interpretation).

There is, in other words, *in effect* only two very roughly outlined positions which are widely discussed by philosophers with respect to the metaphysics of credences and utilities—and consequently, there is some tendency to reject characterisational representationism out of hand, as belonging to an outdated (behaviourist) or *prima facie* implausible (anti-realist) point of view. As we will see, though, there is nothing inherently anti-realist or behaviouristic about the application of representation theorems to the characterisation of credences and utilities.

4.1 Minimal realism and psychological reality

As noted in §2.1, I will assume a *minimal realism* regarding credences and utilities. By 'minimal realism', I mean that ordinary agents in ordinary circumstances have, as an objective matter of fact, credences and utilities; and furthermore, our talk of credences and utilities is not a mere *façon de parler* for talk about outright beliefs and desires. I will therefore set aside any kind of *eliminativism* about the graded attitudes: these states *exist*, and whatever it may turn out to be for S to have a credence of x in P (or a utility of y in Q), it will amount to something other than just being in some outright belief (or desire) state. Minimal realism may well turn out to be false, but I will not address that possibility here.

A more pertinent distinction for our purposes concerns the *psychological reality* of credences and utilities. Call a state *psychologically real* just in case it can be uniquely identified with some natural kind found at a lower level (e.g., computational, neurobiological, etc.) psychological description of the mind. So, for instance, the state of *being in pain* is psychologically real if it can be uniquely identified with some interesting neurobiological state (e.g., c-fibres firing) or basic processes-level state (e.g., a unique computational role) shared by all and only those in pain, where the neurobiological or process-level kind in question can be specified independently of the ordinary causal properties typically associated with pain. If there were nothing inside the head which unified all typical human subjects who are *in pain* beyond the fact that they tend to say ‘ouch’ and search for painkillers (etc.), then *being in pain* would only be surface deep: no interesting part or process involved in the causal workings of the brain would correspond uniquely to pain, so it wouldn’t be psychologically real.

A common view is that the *outright* propositional attitude are psychologically real—or, at least, directly and systematically grounded in something psychologically real. The corresponding view for graded propositional attitudes is likely to be roughly as pervasive. Let us refer to this as *psychological realism* about credences and utilities.[34] Psychological realism can be contrasted with two distinct positions, which are not to be confused: *psychological non-realism* and *psychological anti-realism*. The latter (anti-realism) is the view that credences and utilities are neither psychologically real nor directly and systematically grounded in some psychologically underlying state. By contrast, non-realists positions are designed to be neutral regarding the issue of psychological reality.

Because psychological non-realism is historically more closely associated with characterisational representationism, we will begin our discussion with them; then, in §§4.3–5, we will consider the possibility of developing a realist characterisational representationism.

4.2 Two kinds of non-realism

There is a long-standing non-realist approach of beliefs and desires which shares an obvious resemblance to the kinds of characterisational representationist positions discussed in §3.2. It is generally linked to the following passage in (Ramsey 1927):

> It is, for instance, possible to say that a chicken believes a certain sort of caterpillar to be poisonous, and mean by that merely that it abstains from eating such caterpillars on account of unpleasant experiences connected with them. The mental factors in such a belief would be parts of the chicken’s behaviour … it might well be held that in regard to this kind of belief the pragmatist view was correct, i.e. that the relation between the chicken’s behaviour and [the state of affairs which form the content of the belief] was that the actions were such as to be useful if, and only if, the caterpillars were actually poisonous. Thus any set of actions for whose utility P is a necessary and sufficient condition might be called a belief that P… (144)

Let’s use *pragmatism* for the kind of view being expressed here. Besides Ramsey, pragmatists include Braithwaite (1946), Marcus (1990), and on some readings, Dennett (1971,

[34] For ease of reading, I will usually neglect to specify whether I am talking about psychological realism *about* beliefs and desires or *about* credences and utilities. This should be clear from context.

1989, 1991) comes at least very close to pragmatism. Pragmatism is also discussed (but not endorsed) in (Stalnaker 1984, 1-17) and (Joyce 1999, 19-22). The basic idea behind pragmatism is that to believe that *P* and to desire that *Q* is (*ceterus paribus*) to behave, or be disposed to behave, in such a way as would tend to make it the case that *Q* were it the case that *P* (and all your other beliefs) were true. It will be helpful to refer to this as the *Belief-Desire Law*:

> **Belief-Desire Law**
> If *S* believes that *P* and desires that *Q*, then (*ceterus paribus*) *S* will (be disposed to) behave in such a manner as would tend to bring it about that *Q* if *P* (and all of *S*'s other beliefs) were true

On the pragmatist's approach, the Belief-Desire Law is not a mere empirical hypothesis specifying some contingent regularity which may or may not turn out to be true. Instead, the law plays a *constitutive* or *definitional* role: to be an agent—to have beliefs and desires at all—*is* to (be disposed to) behave in such a manner as would make sense under a given assignment of beliefs and desires under the assumption of the Belief-Desire Law.

One of the central recurrent complaints about pragmatism is that our behaviour usually seems compatible with multiple, inconsistent interpretations:

> What makes an assignment of a system of belief and desire to a subject correct cannot just be that his behaviour and behavioural dispositions fit it by serving the assigned desires according to the assigned beliefs. The problem is that fit is too easy … Start with a reasonable [system of beliefs and desires], the one that is in fact correct; twist the system of belief so that the subject's alleged [beliefs] is some gruesome gerrymander; twist the system of desire in a countervailing way; and the subject's behaviour will fit with perverse and incorrect assignment exactly as well as it fits the reasonable and correct one. (Lewis 1986, 38, see also Stalnaker 1984, 17-18).

There are, of course, some who are willing to bite the bullet of radically underdetermined beliefs and desires, but it's a big bullet to bite. The more common response is to supplement the view with some further principle, which can be used to narrow down the range of available interpretations. We will return to this idea shortly.

Pragmatism—here a view about outright beliefs and desires—is in spirit very close to the kinds of position discussed in §3.2. Indeed, many will want to treat the principle of expected utility maximisation as an *explication* of the folk Belief-Desire Law (just as numerically represented credences and utilities can be taken to explicate the folk notions of belief and desire), and likewise take a representation theorem to underwrite a more precise version of classical pragmatist ideas—one which may even *demonstrably* avoid the underdeterminiation problems, if the theorem's uniqueness conditions are strong enough. After all, the Belief-Desire Law essentially tells us that people generally behave in a manner appropriate to bringing about what they desire given the way they think the world is, and in outline this is what the principle of expected utility maximisation says, albeit in a slightly more refined manner.

Consider, for example, the principle that David Lewis refers to as *Rationalisation*, which forms a central part of his (1974) approach to naturalising intentionality:

> **Rationalisation**
> [A subject] should be represented as a rational agent; the belief and desires ascribed to him … should be such as to provide good reasons for his behaviour, as given in physical terms […] I would hope to spell this out in decision-theoretic terms, as follows. Take a suitable set of mutually exclusive and jointly exhaustive propositions about [the subject's] *behaviour* at any given time; of these alternatives, the one that comes true according [the physical facts] should be the one (or: one of the ones) with *maximum expected utility* according to the total system of beliefs and desires ascribed to [the subject] at that time… (1974, 337, emphasis added)

The basis for Rationalisation, according to Lewis, *is* folk psychology:

> Decision theory (at least, if we omit the frills) is not esoteric science … Rather, it is a systematic exposition of the consequences of certain well-chosen platitudes about belief, desire, preference, and choice. It is the very core of our common-sense theory of persons, dissected out and elegantly systematized. (1974, 337-8)

It is worth mentioning that a commitment to interpreting agents as expected utility maximisers does *not* imply a commitment to interpreting agents as having probabilistically coherent degrees of belief. While something like expected utility maximisation may be involved in the folk theory of the mind and intentional behaviour, it's not so clear that the folk conceptualise ordinary agents as being probabilistically coherent—this is a further commitment of CEU.

It would, of course, be entirely natural to cash out Rationalisation using an expected utility representation theorem suited *in particular* to a behavioural construal of $\succcurlyeq$, and it's plausible that Lewis had some such theorem in mind when he wrote the above passages. After all, many well-known representation theorems (such as Savage's) purport to take us from an agent's behavioural preferences to a unique model of that agent as an expected utility maximiser, which is *exactly* the kind of thing that an interpretive principle like Rationalisation seems to require.

However, the full account that Lewis lays out in his (1974) does not rely *solely* on the principle of Rationalisation—and as a result of this, while it fits naturally with characterisational representationism, it does not imply CCR.[35] Instead, according to what we will call *Lewisian interpretivism* (as opposed to pragmatism), the correct interpretation of an agent is the one (or ones) which maximise fit both with the principle of Rationalisation and also a second interpretive principle, Charity.[36] Charity principles assert that any assignment of doxastic states to an agent ought to maximise some epistemic good(s), such as knowledge, justification, truth, or accuracy. As Lewis understood it, a subject ought to be represented as "believing what he ought to believe" according to what he described as a "common inductive method *M*":

> There must exist some common inductive method M which would lead to approximately our present systems of belief if given our life histories of evidence, and which would likewise lead to approximately the present system of beliefs ascribed to [the subject] if given [the subject's] life history of evidence according to [a purely physical description of that history]. (1974, 336)

Lewis never specified how he intended to naturalistically characterise a "life history of evidence" in a plausible way, though he seems to have taken it as obvious that it could be described in wholly non-intentional terms (cf. Pautz 2013, 220-6). Perhaps he had in mind a complete physical description of the sequence of outside influences upon the agent's sensory organs, along with a physical description of the workings of those organs.

In any case, to appeal to Charity in one's account of the attitudes is to include information about the agent which goes beyond her (actual or counterfactual) patterns of behavioural preferences. Moreover, Charity and Rationalisation principles can pull in quite different directions. Suppose that S satisfies a theorem T's preference conditions so as to be uniquely representable as an expected utility maximiser with credences $\mathcal{B}el$. However, suppose also that $\mathcal{B}el$ is radically at odds with what we would expect S to believe given her life history of evidence. For instance, $\mathcal{B}el$ may assign a very high degree of belief to P, despite the vast majority of S's evidence pointing towards $\neg P$. In this situation, Charity

[35] Some of Lewis' views regarding mental content are also detailed in his (1975), (1983, 373ff), and (1986, 27-50), and seem to have changed only slightly over the years—most of the changes being due to a growing emphasis on the importance of 'naturalness' considerations. See (Weatherson 2012b), (Pautz 2013, 220-6), and (Schwarz 2014a) for helpful, and generally very sympathetic, exegeses of Lewis' interpretivism as expressed in his (1974).

[36] See Davidson (1973, 1990) and Stalnaker (1984, Ch. 1) for positions which share much in common with Lewis's view as described here. Dennett, especially in his (1989, 17-21), also shows strong concern for something like a principle of Charity to interpretation. Lewis' own understanding of Charity also included principles for the reasonable assignment of desires to a subject; e.g., that one should not be interpreted as having an intrinsic desire for a saucer of mud.

should presumably pull us away from the assigning $\mathcal{B}el$ as the agent's credence function.[37] Given that the correct interpretation is constrained by both Charity and Rationalisation, and so long as Rationalisation is not given any strong interpretive priority over Charity, it would be unreasonable for the Lewisian interpretivist to assign $\mathcal{B}el$ in her final interpretation of *S*. Given an appeal to Rationalisation—which is naturally precisified by means of a representation theorem—Lewisian interpretivism suggests characterisational representationism, but not CCR.

For both pragmatists and interpretivists alike, propositional attitudes might best be thought of as *states of a person* rather than *states of* (or *in*) *the head*: the workings of the brain are irrelevant on both kinds of views, what matters is just one's behavioural states and (for the Lewisian interpretivist) one's history of evidence. Propositional attitude attributions are not hypotheses about the inner workings of the brain; they are instead usually conceptualised as parts of an innate theoretical system (folk psychology) developed over time for the explanation and prediction of behavioural patterns, but where that folk theory involves no strong commitment to psychological realism.

A classic intuition pump in favour of this line of thought goes as follows. Imagine that we are visited by a race of alien beings, whose internal physical constitution is entirely unknown to us, but who are able to speak our languages, engage in intelligent and meaningful conversations and apparently express very sensible thoughts, react as we would to various stimuli, and generally behave just as any ordinary human would across a huge range of contexts. It seems entirely natural to describe such beings as having beliefs and desires (or credences and utilities), and to explain their apparently intentional actions by reference to those beliefs and desires—even in complete ignorance of whatever may be going on inside their heads. Indeed, we may even suppose that there is nothing inside their heads which correspond to an ordinary human brain. Almost certainly, the aliens will have to in some way or another represent how they take the world to be and other ways it might be, but this need not be through any psychological structures akin to human representations. To the extent that we are still willing to assign beliefs and desires to these aliens, the implication seems to be that what ultimately matters for propositional attitude attribution is grounded at least in part in patterns of actual and counterfactual behaviour, independent of any questions concerning the psychological reality of those attitudes.

As we have noted, though, the view that the propositional attitudes must be psychologically real (or directly grounded in some psychologically real state) is very common amongst contemporary philosophers, who are *ipso facto* liable to reject any non-realist account of credences and utilities. In the next few sections, we will consider whether

[37] In saying this, I am of course disagreeing with Maher (see §3.2). Maher appeals to a CEU theorem and asserts that because $\mathcal{B}el$ is probabilistically coherent, it is therefore part of a *perfect* interpretation of *S*. But, probability function or not, $\mathcal{B}el$ is unlikely to accurately represent *S*'s credences if it doesn't take into account her reasonable response to evidence.

representation theorems might be fruitfully applied to the development of a psychologically realist position instead. But, before we move on, it's worth noting that the two versions of non-realism outlined in this section are *entirely* consistent with the kind of functional role semantics that I will suggest in §§4.4–5. Indeed, should psychological realism turn out true, then both pragmatism and Lewisian interpretivism very naturally suggest a functional role semantics for the underlying psychologically real states, cashed out at least in part by means of a representation theorem.

4.3 Psychological realism and the structure of thought

A popular idea amongst psychological realists (or just *realists*, for short) is that the mind works in a manner closely analogous to a digital computer, and that the outright propositional attitudes are (or are very closely connected to) physical *data structures* stored somewhere in the brain—much like files stored on a computer's hard drive, ready to be used should the need arise. These structures are usually referred to in the jargon as 'mental representations', though understood as such 'mental representation' is a technical term used by theorists working within the so-called *computational theory of mind* (see Fodor 1975, Putnam 1980). In what follows, I want to generalise away from specific theories of cognition (e.g., computationalism versus connectionism), so to avoid ambiguity we will use '*M*-representation' to refer to any psychologically real state with semantic properties—whether these semantic properties are truth values, truth-conditions, success conditions, reference, or otherwise.

As a first pass characterisation only, many realists would be happy to assert that for a *non-graded* propositional attitude ϕ, *S* ϕ*s that P* just in case *S* has somewhere in her head an *M*-representation *#P#* that (a) plays a ϕ-like role in cognition, and (b) means that *P*. What exactly constitutes a ϕ*-like role* is never specified precisely—instead we usually find a promissory note that such details will be fleshed out eventually (in fully naturalistic terms no less). Presumably, a belief-like role, for example, would involve certain patterns of responses to perceptual states, being a guide in action when taken in combination with one's desires (as per the Belief-Desire Law), and so on. I will discuss this further below.

It is widely held that this first pass characterisation is too strong. Consider, for example, the case of belief. Most realists will hold that we have many implicitly held beliefs—such as the belief that *1000 is less than 1001*, that *1001 is less than 1002*, and so on—without holding that we have for *every* such belief some *M*-representation with the content of that belief (and *only* that content) stored somewhere in the head. To think otherwise would be to countenance a massive (if not infinite) proliferation of stored informational structures, which would be psychologically implausible.

The most common response to such considerations is to weaken the earlier characterisation, giving us what I will call *Basic Psychological Realism* (or BPR):

> **Basic Psychological Realism (BPR)**
> *S ϕs that P* iff *S* has in her head an *M*-representation *#P#* that (a) plays a *ϕ*-like functional role, and either (b) means that *P* or (c) has some other content from which *P* can be 'readily extracted'

As will become clear, what the "other content" may be, and what "readily extracted" means, will depend on the specifics of the view to be developed. In the event that (a) and (b) hold, we can say that *S explicitly ϕs that P*, whereas *S implicitly ϕs that P* when *only* (a) and (c) hold. In another manner of speaking, *explicit* attitudes are psychologically real states, whereas *implicit* attitudes are directly grounded in psychologically real states. (Thus, the psychological realist about beliefs will hold that *all* of our beliefs—whether explicit or implicit—either are, or are directly and systematically grounded in, psychologically real states.)

If BPR is true, then there are a number of options regarding what to say about the exact character of the *M*-representations that underlie our propositional attitudes, and the origins of their content. According to BPR, if *S ϕs that P* then there must be some *M*-representation (call it *#P#*) which either means that *P* or is otherwise closely connected to *P* by virtue of whatever content it does have. In what follows, I will first characterise two prominent philosophical positions on the *structure* of these underlying *M*-representations, before turning to a discussion of how they might get their content in §4.4. These two positions were developed with outright beliefs and desires in mind; as we proceed, I will also discuss how they might also be augmented to apply to the graded attitudes.

Perhaps the most common—or at least the most commonly discussed—view on the character of *#P#* originates with Fodor (1975, 1987), who held that:

(i) We have many distinct *explicit* beliefs and desires

(ii) The *M*-representations *#P#* underlying each explicit attitude state have propositional content

(iii) *#P#* has an internal structure that's closely analogous to the sentences in spoken languages used to express those contents

Each of these is a non-trivial, empirical claim, and the conjunction of all three we can refer to as the *sentential* view. The third claim is particularly important; the idea is that, just as sentences expressing propositions are constructed out of words with sub-propositional contents, so too might we think that the *M*-representations directly underlying our beliefs can be broken down into more basic *M*-representations—called *concepts*—with stable sub-propositional contents, which have to be composed in the right way to arrive at the right proposition. So, for example, on the sentential view the *M*-representation which means that *John is taller than Frank* might be composed out of the concepts *#taller#*, *#John#*, and *#Frank#*, and have the structure <*#taller#*: <*#John#*, *#Frank#*≫, where reversing the order of *#John#*, and *#Frank#* would alter the proposition thereby expressed.

A central motivation for the sentential view is its capacity to explain the systematicity and productivity of thought. Thought is *systematic* in the sense that the ability to entertain some contents seems to come hand-in-hand with the ability to entertain others. The ability to believe that (or desire that, etc.) *John is taller than Frank* seems to imply also the ability to believe that (or desire that, etc.) *Frank is taller than John*. Thought is also *productive* in the sense that we seem to have the ability to entertain an unlimited number of contents; the beliefs I actually have, for example, are just a fraction of the beliefs that I *could* have had. Sentential views explain these two features of thought by positing a range of stored concepts with fixed contents which can be freely recombined in an unlimited number of ways according to simple rules to produce an unlimited number of sentence-like *M*-representations with distinct propositional contents.

Advocates of sentential views like to speak of 'belief boxes' and 'desire boxes' as metaphors for the set of stored *M*-representations which play belief-like and desire-like roles respectively. When tasked to say whether she believes that *P*, a subject is conceived of as searching through the sentences contained in her belief box to find one which either reads *P* or ¬*P*, or in lieu of that, some other sentence from which either *P* or ¬*P* readily follows. Importantly, on this picture, an ordinary subject is generally seen to have a great many *explicit* beliefs and desires: there are many sentences stored in her belief and desire boxes, and those sentences are about reasonably non-specific matters—about as specific as an ordinary assertion in a natural language (e.g., 'Roses are red' or 'Australia has 6 states').

Sentential views appear to be the most prominent view amongst philosophers who presuppose the psychological realism of the outright propositional attitudes. Unfortunately, they are usually discussed in the context of debates surrounding *outright* beliefs and desires, and the extent to which they might be applied to credences and utilities does not seem to have been anywhere thoroughly explored—although I suspect that many of the philosophers who incline towards a sentential view for beliefs and desires would hold a similar view for credences and utilities (to the extent that they take the latter to be psychologically real).[38]

[38] In a recent paper, Goodman *et al*. (2015) outline what they call the 'probabilistic language of thought hypothesis', that "concepts have a language-like compositionality and encode probabilistic knowledge" (626). Their account is similar to Fodor's in that the *M*-representations underlying our credences are assumed to be structures in a computational system built out of recombinable parts with stable contents in a broadly language-like fashion (claim (iii), above). However, instead of positing sentence-like *M*-representations with propositional contents (one for each explicit credence state), Goodman *et al*.'s 'sentences' jointly encode probability distributions over a space of possible world states. In this respect, their position shares more in common with the map-like views discussed below.

Let us consider how one might incorporate credences into a sentential account. There are two obvious options here; the first option is to suppose that every relevant sentence-like *M*-representation comes with some attached psychological property that corresponds to a degree of confidence *x*, which determines for it a unique cognitive role (i.e., a 'credence-of-*x*-like' role). If so, it's easy enough to extend BPR to account for *explicit* credence states:

> *S has an explicit credence of x in P* iff *S* has an *M*-representation *#P#* that (a) plays a credence-of-*x*-like role, and (b) means that *P*

Instead of a 'belief box', a better metaphor here would be a 'credence warehouse': picture a large warehouse containing a number of equal-sized barrels, each labelled with a unique sentence and each containing some amount of 'confidence fluid'. One then has an explicit credence of *n*/100 in *P* if, somewhere in this warehouse, there is a barrel labelled with a sentence that means that *P* that is *n*% filled with confidence fluid.

This account already implies a large proliferation of stored *M*-representations—far more so than were needed to account for our explicit beliefs, for now we need to account for a range of explicit credences towards a huge range of possible degrees of confidence. Moreover, it's not obvious how we might make sense of *implicit* credence states under this kind of view. When it comes to outright beliefs, it's somewhat plausible that if *S* explicitly believes that *P* and *S* can derive *Q* from *P* with very little effort, then *S* also believes that *Q* (albeit only implicitly). However, the strategy of appealing to "easy derivations" does not seem to apply where it is credences rather than beliefs that are the focus of the account. Relations between credence states function differently than relations between outright beliefs—credences operate under a different logic. If *S* has an explicit credence of *x* in *P*, and *Q* can be readily derived from *P*, what should be said about *S*'s credence towards *Q*? Perhaps *S* does have *some* implicit degree of confidence towards *Q* in this case—but what degree? Probabilists will assert that *S*'s credence in *Q should be no less* than *x*, but this is quite uninformative even under the (implausible) assumption that *S* is probabilistically coherent. Indeed violations of monotonicity are empirically well-established, even in cases where the relevant *Q* is easily derivable from *P* (see the conjunction fallacy discussed in Tversky and Kahneman 1974).

There is another option for sentential views, which is to appeal to a position that we might call *ordinalism*. According to ordinalism, *absolute* credence states reduce to *relative* credence states (see §2.1 for the absolute/relative distinction). On this picture, the degrees of confidence assigned to individual propositions are a measure of their position within an overall *ordering* of claims according to their relative plausibility. A sententialist picture fits nicely with ordinalism, supposing at least that the ordering can be appropriately functionally characterised. In particular, the idea would be that the psychological reality of our credence states (both absolute and relative) is to be explained by positing a large collection of sentences arranged (in an appropriate functional sense) according to their relative perceived likelihoods. Then, to find P more plausible than Q would be to have a sentence which means that P situated higher in the ordering than a sentence that means that Q. On the other hand, to have a credence of x in P would be to have a sentence which means that P whose location relative to other sentences in the ordering is represented by the degree x. Sentences at the bottom of the ordering are conventionally assigned a value of 0, those at the top are assigned a value of 1, and everything else gets assigned a value in between.

This position will again face the issue of avoiding an excessive proliferation of stored M-representations—it certainly seems implausible that for every proposition P towards which we have some credence there must be a sentence #*P*# sitting somewhere in the ordering, so we would need an appropriate way of understanding implicit credence states. But there is another challenge here. It is generally assumed that credences are not merely measured on an *ordinal* scale, that a credence function should be understood at least as an *interval* scale, if not a *ratio* scale. *Intervals matter*: if S believes P to degree 0.1, Q to degree 0.2, and R to degree 0.4, then the difference in her degree of belief between P and Q is less than the difference between Q and R. Indeed, it even seems that *ratios matter*: it's natural to say that I am *twice* as confident that a fair coin will land heads if it's flipped once (credence 0.5) than I am that it will land heads twice in a row (credence 0.25). In order to show that an ordering $\succcurlyeq^{\mathrm{x}}$ of some collection of entities—in this case, a collection of sentence-like M-representations—can be measured on either an interval or a ratio scale, we need to show that $\succcurlyeq^{\mathrm{x}}$ satisfies a number of further structural constraints. (In particular, we need to *at least* be able to say when *the difference in degree between* #*P*# *and* #*Q*# is equal to *the difference in degree between* #*R*# *and* #*S*#, and show that this quaternary relation satisfies certain structural conditions—see Definition 8.6.) The challenge, then, is to establish an empirically plausible set of constraints on a functionally-characterised ordering of M-representations which will allow for an appropriate measure of credence—i.e., a measure which goes beyond merely ordinal information, which lets us represent the relative *strengths* with which propositions are believed.[39]

Sentential views are not universally accepted. Scepticism regarding the approach is famously associated with Dennett (1971, 1989, 1991) and Stalnaker (see esp. his 1976, 1984, 1999b). An important alternative to understanding the character of #*P*# is to treat it as having a structure and content analogous to that of a street map. (See Lewis 1982, 1994, Braddon-Mitchell and Jackson 1996, Ch. 10.) Call this a *map-like view*. A map is

[39] It would, I think, be a mistake to appeal to a representation theorem for a system of qualitative probability (mentioned in §2.4) in spelling out this position, as opposed to one of the more traditional theorems for extensive measurement (see Krantz, Luce *et al.* 1971). Theorems like de Finetti's (1931), which allows us to T-represent a weak ordering $\succcurlyeq^{\mathrm{b}}$ over an algebra of propositions $\mathcal{P}$ using a probability function, rely heavily on set-theoretic relations between the propositions in $\mathcal{P}$—relations which we shouldn't assume hold between the sentences used to express those propositions.

not a collection of sentences, nor does it hold information in the same way that a collection of sentences does (Camp 2007). An ordinary street map, for example, is a single informationally-rich representational object which, due to the arrangement of its parts, manages to hold information about the relative position, orientation, number, dimensions and names of a large number of distinct entities (buildings, streets, hills, etc.).

The hypothesis that *M*-representations might be more akin to maps than sentences is intended to help explain the productivity and systematicity of thought (and thus serve as a counterexample to the claim, sometimes made, that these two properties can *only* be explained given a sentential view). Maps are, for one thing, *systematic*: the way a map represents one aspect of the world is closely integrated with how it represents a great many other things. For instance, one cannot change the absolute location of a hospital on a map without also changing its position relative to everything else, changing the shortest path to the hospital from a given location, and so on. Likewise, if *M*-representations are map-like, then thought is also plausibly *productive*: an alteration in the arrangement of the parts of a map, or the addition of a new part, produces a new representation of the way the world is. By analogy, a map-like *M*-representation is supposed to be a single, highly integrated and informationally rich representational unit which captures information about a great many things at once, where a small change in its structure might mean a great many changes in the specific informational content that it holds.

In contrast to the sentential view, proponents of map-like views seem to prefer the idea that there might be relatively few *M*-representations underlying our beliefs and desires—perhaps even only *one* for each kind of attitude. As Lewis puts it,

> If mental representation is map-like … then 'beliefs' is a bogus plural. You have believes the way you have the blues, or the mumps, or the shivers. But if mental representation is [sentence]-like, one belief is one sentence written in the belief box, so 'beliefs' is a genuine plural. (1994, 311)

Our individual beliefs, such as the belief that *roses are red* or that *Tuesdays follow Mondays*, are conceived of as different fragments of information extracted from a single, complex and highly-structured *M*-representation, which encodes our *overall* picture of how the world is. This is consistent with BPR, though it implies that what we would usually call our 'beliefs' are, in general, *implicit* beliefs—to the extent that there are any psychologically real doxastic states, they corresponds to whole *systems of belief*, rather than to *individual beliefs*.

Lewis also argues for another difference between map-like and sentence-like *M*-representations: "Mental representation is [sentence]-like to the extent that parts of the content are the content of parts of the [*M*-]representation", whereas "If our beliefs are [like maps], then they are to that extent not language-like" (1994, 310-11). As Blumson (2012) points out, though, parts of a map often represent parts of what the map as a whole represents—the bottom half of a map of the Earth usually represents the geography of the southern hemisphere, for instance. I suspect, however, that Lewis was casting doubt on the idea that our beliefs are *nothing more than* structures composed wholly out of a finite base of discrete and freely recombinable elements (i.e., concepts) with fixed contents—that is a commitment of the sentential view, and while it's consistent with the map analogy, it should not be taken for granted. Mental representation need not be *digital*, and the parts of an *M*-representation (to whatever extent they can be isolated) need not have significance independent of their role within a broader context (cf. Camp 2015).

Map-like views are relatively underdeveloped, and there does not seem to have been much of an attempt within philosophy to extend map-like views to deal with credences and utilities—though recent work on causal Bayes nets in psychology could be of much use here. See, especially, (Pearl 1988, 1990) and (Gopnik, Glymour *et al.* 2004). The discussions in (Lewis 1982, 1994) and (Braddon-Mitchell and Jackson 1996) focus on how a map-like view might work as an account of our *beliefs*: according to these authors, map-like *M*-representations are taken to capture an entire belief system by virtue of representing a single, highly specific way the agent believes the world to be. The map essentially picks out a set of *doxastically possible worlds*—a highly specific proposition—and the agent is said to believe any proposition *P* which is true at every such world.

However, if a single *M*-representation is to underlie all of our credences, then it clearly must take a quite different structure than that of a map which merely represents just *one* way that things might be. In particular, it needs to be able to represent a very wide range of ways things might be, along with their respective likelihoods. Instead of an ordinary street map, then, which represents one way things might be, perhaps what would be needed is a more complicated 'map' of some space of possibilities, with different areas of the map being marked as more or less likely. The content of the *M*-representation, in other words, might have more in common with a *probability density distribution* over a space of possibilities than it does with a street map that represents a single way the world might be (see Figure 4.1). This, at any rate, seems to be how Lewis (1986, 30) imagines an extension of a map-like view to account for credences.

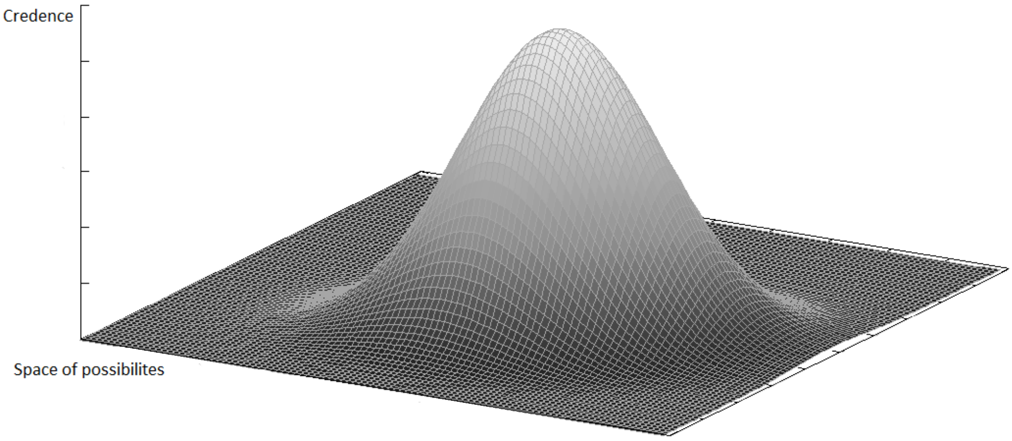

Figure 4.1

To flesh out this idea more, let $\mathcal{W}$ be a space of possibilities. We will suppose that $\mathcal{W}$ is finite, but we need not suppose that the elements of $\mathcal{W}$ are maximally specific—it's enough that they are mutually exclusive and jointly exhaustive; i.e., the actual world must be in exactly one $w \in \mathcal{W}$. A *probability density distribution* on $\mathcal{W}$ is a function $\mathcal{D}$ which assigns each $w \in \mathcal{W}$ a real value somewhere between 0 and 1 such that:

$$\sum_{w \in \mathcal{W}} \mathcal{D}(w) = 1$$

$\mathcal{D}$ effectively assigns a credence value to every element of $\mathcal{W}$. Furthermore, we can use it induce a probability function $\mathcal{P}r$ on an algebra of sets $\mathcal{P}$ constructed from $\mathcal{W}$, by simply assuming that for each $P \in \mathcal{P}$,

$$\mathcal{P}r(P) = \sum_{w \in P} \mathcal{D}(w)$$

Thus, a single *M*-representation which encodes something like a probability density distribution over a space of possibilities can *ipso facto* also be taken to encode an agent's credences towards any propositions which can be constructed out of that space. There is, however, an immediate problem with this way of understanding map-like *M*-representations for credences: given that I have characterised $\mathcal{W}$ as a space of *possibilities*, $\mathcal{D}$ can only be taken to encode the credences of a fully probabilistically coherent subject.

The beginnings of a solution to this problem might be found if $\mathcal{W}$ is allowed to include *im*possible states of affairs, to which $\mathcal{D}$ might assign values of greater than 0. Under this kind of construction, different impossible propositions can be modelled as distinct regions within $\mathcal{W}$ and can be assigned some positive credence. Likewise, distinct necessary propositions will be distinct regions within $\mathcal{W}$—they will intersect with respect to the possible states of affairs, but will differ at various impossible states of affairs. Because necessary propositions are no longer just equivalent to the universal set $\mathcal{W}$, they may be assigned values of less than 1.

Finally, and perhaps most interestingly, the probability function $\mathcal{Pr}$ induced by $\mathcal{D}$ need not be 'additive', *in the sense that* if P logically implies $\neg Q$ and Q logically implies $\neg P$, then:

$$\mathcal{Pr}(P \vee Q) = \mathcal{Pr}(P) + \mathcal{Pr}(Q)$$

That is, suppose that $P \vdash \neg Q$ and $Q \vdash \neg P$; i.e., there are no *possibilities* in $\mathcal{W}$ where both P and Q hold. However, there may still be *impossibilities* in $\mathcal{W}$ where P & Q is true. Let w be any such impossibility. If $(w) > 0$, it immediately follows that:

$$\mathcal{Pr}(P \vee Q) < \mathcal{Pr}(P) + \mathcal{Pr}(Q)$$

This shows that the representation of a credence state by means of a probability function is *consistent* with subadditive credences, if the function's domain is adequately structured.[40] Using similar reasoning, it's also easy to show that $\mathcal{Pr}(P)$ may be greater than $\mathcal{Pr}(Q)$ even when $P \vdash Q$.

With the right kind of impossible states of affairs in $\mathcal{W}$, $\mathcal{Pr}$ may also assign superadditive credences.[41] In particular, suppose that there is an impossibility w in $\mathcal{W}$ where $P \vee Q$ is true but neither P nor Q is true, for a pair of logically incompatible P and Q, and suppose that $(w) > 0$. Then, assuming that no credence is given to any impossibilities where P & Q, it will turn out that:

$$\mathcal{Pr}(P \vee Q) > \mathcal{Pr}(P) + \mathcal{Pr}(Q)$$

Thus, a map-like M-representation might encode something like a probability density distribution over a possibility space (broadly construed), and thereby also encode even highly incoherent credences over a wide range of propositions more generally. But there are also other options here—for instance, Dubois and Prade's (1988) *possibility theory* allows us to systematically construct a credence function on the basis of what they call a *possibility distribution*; i.e., a function $\mathcal{D}'\colon \mathcal{W} \mapsto [0, 1]$ such that $\mathcal{D}'(w) = 1$ for at least one

[40] Interestingly, this kind of 'sub-additivity' arises precisely because $\mathcal{Pr}$ is *additive*, in the technical sense that for all $P, Q \in \mathcal{P}$, if $P \cap Q = \emptyset$, then $\mathcal{Pr}(P \cup Q) = \mathcal{Pr}(P) + \mathcal{Pr}(Q)$. On this construal of $\mathcal{W}$, it is no longer true that $P \vdash Q$ implies $P \subseteq Q$ (though the reverse still holds).

[41] Thanks to Daniel Nolan for reminding me that this can also be done. There are problems that arise with trying to systematically characterise the relevant impossibilities; see (Bjerring 2013).

$w \in \mathcal{W}$.[42] Likewise, appeal may be made to the Dempster-Schafer theory of *belief functions*, wherein each proposition P divides a possibility space into three sections—one where P is determinately true, one where P is determinately false, and one where it's unclear whether P or $\neg P$—which seems especially useful in capturing imprecise credence states (see Dempster 1968, Shafer 2011).

Let us summarise. The key point of the foregoing discussion is that to be a psychological realist about credences (and likewise for utilities), one does not have to hold that there is any straightforward, *one-to-one* correspondence between the absolute credence states we attribute to ourselves and others, and whatever M-representations might be found inside the head. Theorists who adhere to a map-like view about *beliefs* hold that those beliefs really are *in the head*—i.e., there is something psychologically real which directly and systematically grounds the truth of our belief attributions—but they don't think that having many individual beliefs with such-and-such contents requires having many M-representations with just those contents. Psychological realism, as it is here being understood, is not equivalent to the sentential view.

4.4 Content determination for realists

Suppose that S has a credence of x in P. (Alternatively, suppose that S has a utility of y in P—the focus on credences rather than utilities here is immaterial to the discussion.) The Basic Psychological Realist is then committed to holding that there is some M-representation *#P#*—whether sentence-like or map-like or otherwise—which is the psychological basis for S's credence, and that *#P#* has some content or other. What grounds *#P#*'s content, whatever that content may be?

In what follows, I will look at different strategies for answering this question, and argue that representation theorems seem especially useful in fleshing out the details for one of these strategies in particular (what will later be referred to as *functional role semantics*).

One common strategy we might call *compositionalist*: *#P#* is to be broken down into smaller, independent M-representations with fixed contents—i.e., concepts—and, following an appropriate account of conceptual content, we are to work out *#P#*'s content using some principle of compositionality. According to the compositionalist, intentionality first enters the mind through our concepts, not through our propositional attitudes—propositional content is derivative upon conceptual content. Partly because of the prominence of sentential views, the compositionalist strategy has proven especially common over the past few decades. If each of our attitudes involves a sentence-like *#P#* playing a ϕ-like

[42] In particular, for any non-empty P in an algebra of sets on $\mathcal{W}$, define the *possibility measure* f as follows: $f(P) = sup\{\mathcal{D}'(w) \colon w \in P\}$, and $f(\emptyset) = 0$. This implies that f is a sub-additive credence function: $f(P \cup Q) = max\{f(P), f(Q)\}$, while $f(\mathcal{W}) = 1$.

role in cognition, then a natural way to approach the project of characterising what it is to ϕ *that P* would be to try to break the task into three (presumed to be independent) subtasks, to be approached in the following order:

(i) Account for the content of the constituents of *#P#*
(ii) Explain how those parts compose to produce *#P#* and its propositional content
(iii) Explain what it is for *#P#* to play a ϕ-like role

For instance, suppose that *S* (explicitly) believes that *cats are mammals* and *cats are friendly*, and (explicitly) desires that she possess a cat. Tokens of the concept *#cat#* appear as constituents in each of these attitudes—as Fodor puts it, "some mental formulas have mental formulas as parts; and … those parts are 'transportable': the same parts can occur in many mental representations" (1987, 137). Implicit in this is that the meaning of *#cat#* is stable across 'transportations'. It is natural, then, to think that the explanation for why *#cat#* refers to *cats* should be the same in each of these instances, whatever that explanation might be. If we can find that explanation and apply it to all of our concepts, then *most* of the hard work in explaining the propositional attitudes will be complete—the only remaining tasks would be to show how concepts can be used to build sentence-like structures, and then to differentiate the attitudes by the roles that these structures might play.

There are two very well-known kinds of views on how we might (naturalistically) account for conceptual content that fit nicely with the compositionalist strategy. The first are *causal-informational views*, which appeal primarily to co-variation, or *indication*, relations between concepts and their purported contents. At a first pass, the idea is that *#cat#* means *cat* just in case *#cat#* tends to be tokened in the presence of cats—the tokening of *#cat# indicates* the presence of a cat. More complicated versions of the view might appeal to special formation periods during which the content is fixed and remains unchanged thereafter (Dretske 1981), or asymmetric dependency relations between concept tokenings and potential contents (Fodor 1987).

The second kind of view, *teleosemantics*, can be thought of as a special case of the causal-informational approach—one which appeals in particular to covariation relations under conditions of *proper* functionality, where a concept's *proper function* is understood in biological terms:

> *F* is the (or a) proper function of a characteristic *C* in an entity *E* iff the (or a) reason *E* has *C* is because it *F*s

In the biological case, where *E* is an organism, *F* is *C*'s proper function just in case the possession of *C* conferred a fitness advantage to *E*'s ancestors either wholly or partly because it *F*s. For example, the proper function of a human eye is to see: the possession of eyes conferred a fitness advantage to human ancestors in their normal environments precisely because they allowed them to see; so the fact that eyes enabled sight in those environments is the ultimate explanation for why we have eyes today.

Very roughly, then, a teleosemantic account of conceptual content would be that *#cat#* means *cat* just in case it would be reliably tokened in the presence of cats (and only in the presence of cats), were it to be functioning properly in ancestrally normal environments. (For more details, see Millikan 1989, 1990, Neander 2006.) The concept *#cat#* is, in other words, an adaptation which exists today because it *was* tokened in the presence of cats, and by virtue of this (according to the teleosemanticist) it manages to *now* represent *cats*. (Note that teleosemantic views need not be, and have not always been, tied to sentential views nor the compositionalist strategy. I will return to this below.)

Both of these kinds of views suffer from unresolved problems recognised more or less since their inception. It would take us too far afield to discuss these in any depth, so I will highlight just a couple. Because both causal-informational and teleosemantic views attempt to explain conceptual content ultimately in terms of causal relationships between concepts and the external world, they suffer from so-called *disjunction problems*: when a concept is causally connected to many distinct features of the world—some of which may be co-extensive—it can be difficult (if not impossible) to distinguish one causal relation as *the* relation that is important for fixing the content (see Fodor 1984). A related issue concerns the content of concepts about non-existent objects and uninstantiated properties: such things cannot enter into causal relationships of any kind, and so pose problems for theories that rely on such relationships (see Stampe 1977).

There are alternatives to the compositionalist strategy. For example, consider *inferential role semantics*, expressed here by Paul Boghossian:

> Let's suppose that we think in a language of thought and that there are causal facts of the following form: the appearance in *S*'s belief box of a sentence R_1 has a tendency to cause the appearance therein of a sentence R_3 but not R_2 … we may describe this sort of fact as consisting in *S*'s disposition to *infer* from R_1 to R_2, but not to R_3. Let's call the totality of the inferences to which a sentence is capable of contributing, its *total inferential role* … Against this rough and ready background, an *inferential role semantics* is just the view that there is some construct out of an expression's total inferential role that constitutes its meaning what it does. Let us call this construct an expression's *meaning-constituting inferential role* … (1993, 73-4)

Setting aside some minor complications, the general idea is that R_1 should be assigned a content which best rationalises its meaning-constituting inferential role—essentially, it involves a principle of Charity applied to specific patterns of inference. So, for example, given "*S*'s disposition to infer from R_1 to R_2, but not to R_3", it might be appropriate to let R_1 mean *there are cats* if R_2 means *there are animals* and R_3 means *there are cups*, but it would not be appropriate to interpret R_1 as such if R_2 meant *there are no animals*.

Although Boghossian casts the view in terms of a language of thought, there is nothing in the inferential role semanticist's view of *content determination* requires specifically *sentence*-like *M*-representations—R_1 through to R_3 could be wholly *unstructured* symbols, incapable of being decomposed down into concepts, so long as (a) there are sufficiently many of them, and (b) they can figure in causal relationships with one another and can thereby be assigned propositional contents.[43] Conditions (a) and (b) do suggest that inferential role semantics will not play nicely with a map-like *M*-representations, but map-like views and sentential views aren't jointly exhaustive.

Inferential role semantics resembles the compositionalist strategy in that it attempts to divide the problem of characterising what it is to ϕ *that P* into two (presumed to be independent) sub-tasks, to be approached in the following order:

(i) Account for the content of *#P#*
(ii) Explain what it is for *#P#* to play a ϕ-like role

To the extent that an account of *conceptual* content is then needed, the general strategy is to consider the overall role that concepts play within the sentence-like *M*-representations of which they form a part, and assign a content on that basis—thus giving us a form of *conceptual role semantics* for conceptual content, where conceptual content is taken to be derivative upon propositional attitude content.[44] For instance, we might notice that the concept *#cat#* can be found in all and only the sentences which express something about *cats*, and thereby assign it that content to capture the role it plays in inference (see esp. Block 1986).

Inferential role semantics suffers from general problems relating to the precise specification of the meaning-constituting inferential role. It also has to deal with *permutation problems*, which suggest that *any* specification of a meaning-constituting inferential role might be insufficient for the purposes of pinning down determinate contents (see Lewis 1984, Williams 2007, 2008)—and to whatever extent contents can be pinned down, they

[43] Boghossian's motivation for adopting the language of thought hypothesis is grounded Fodorian considerations regarding the productivity and systematicity of thought, rather than considerations about content determination.

[44] Conceptual role semantics (CRS) comprises a very broad and heterogeneous collection of views, which centre on the idea that the content of a *concept* is determined primarily by the functional role that the concept plays in thought. As it is described here, CRS is not an instance of the compositionalist strategy because the content of a concept cannot be determined prior to fixing the contents of the larger representational structures of which it is or may be a part—CRS treats the compositionalist's sub-tasks (i) and (ii) as highly interdependent. Some versions of CRS may also take into account not only the role that concepts play in licensing inferences between beliefs, but also their connections to perceptual states, categorisation behaviour, and so on, so conceptual content need not be *wholly* determined by propositional attitude content.

may end up being the *wrong* contents (see Williamson 2009). These are well-known problems for an inference-based semantics, but I want to suggest two more issues which seem to me equally serious.

The key idea behind inferential role semantics is that the (propositional) content of the relevant *M*-representations (R_1 to R_3) is supposedly fixed by the causal role that they play, or *would* play, *if* they were deployed in a specifically *belief*-like manner: R_1 to R_3 are symbols which can play many different roles in cognition—giving rise to different kinds of propositional attitude—but it's only the role that they would play were they in a 'belief box' that matters in relation to their content. Where *#P#* underlies a *desire that P*, for example, this is because of what *#P# would* do *were* it to play a belief-like role—it just so happens that *#P#* underlies a desire, though what it *does* in that capacity plays no part in grounding its content.

There is, therefore, an important background assumption made by inferential role semanticists: that the *M*-representations like *#P#* which might underlie *any* given attitude state can always be involved in the relevant kind of inferential relationships—that while *#P#* might *actually* be the psychological basis for a desire (or a utility, or a credence), it *could* be employed in a specifically belief-like manner. (A similar assumption is of course made by advocates of the compositionalist strategy.) That this should be the case is by no means obvious, and neither is the assumption necessary to account for the systematicity and productivity of thought. In particular, a psychological realist could well hold that one's credences regarding *P* and one's utilities towards *P* could be underwritten by two distinct and dedicated types of *M*-representation, *#P#* and **P**, such that it does not even make sense to speak of **P** playing a credence-like (or belief-like, etc.) role.

Moreover, the focus on belief-like inferential relations seems odd, if not simply unmotivated. Consider, for instance, the following (admittedly fanciful) scenario. We note that whenever *S* has in her 'belief box' a sentence-like *#P#*, she is disposed to make certain inferences which, under considerations of Charity, suggest the assignment of content *P* to *#P#*. To this extent inferential role semantics seems on the right track. Suppose, however, that whenever a sentence of the same orthographic type as *#P#* shows up in *S*'s 'desire box', *S* has a strong tendency to reject any course of action which would tend to bring it about that *P*. In other words, *S* acts in a way we would expect if she were to desire that $\neg P$. In this case, to attribute to her a desire that *P* just because *#P#* can be found in her desire box would be absurd. What *#P#* does when it's in *S*'s 'desire box' matters. It's at least conceivable that one and the same kind of *M*-representation might, *by virtue of being involved in different cognitive processes*, underlie (say) a *belief that P* and a *desire that Q*, for very different *P*s and *Q*s.

In other words, it may not be reasonable to disassociate the content of an *M*-representation *#P#* from the particular role that *#P#* plays, and different types of *M*-representation might be tied to particular kinds of attitudes. This possibility leads us, finally, to *functional role semantics*, wherein both the content of an attitude state, and the role played by whatever *#P#* realises it, are accounted for in unison. According to functional role semantics, the meaning of *#P#* and what it does (or should do) are inextricably bound together: if *#P#* is the psychologically real basis for *S*'s *ϕ-ing that P*, then what it does (should do) in that capacity also grounds the fact that it's an attitude *about P*—the connection that *#P#* has with *P* is a function of its unique causal role. Although functional role semantics is not committed to treating *M*-representations as either map-like or sentence-like (or otherwise), it's uniquely well-suited for fixing the content of map-like structures. This is in contrast with the compositionalist and inferentialist strategies, which work best with a sentential view.

It should come as no surprise by now that wherever something like functional role semantics is discussed, there is a very strong focus on characterising (individual) beliefs in particular, though desires occasionally receive some attention. Other kinds of attitudes—including credences and utilities—are rarely mentioned. Regarding beliefs, we generally hear that *#P#* is the basis for a *belief that P* just in case it satisfies all or at least most of the following conditions:

(1) In conjunction with a desire that *Q*, *#P#* leads to behaviour which would tend to bring it about that *Q* at worlds where *P* is true (along with all the subject's other beliefs)

(2) In conversation, *#P#* leads to an assertion that *P* whenever the question of whether *P* or ¬*P* is conversationally salient and sincere assertion is rewarded

(3) Where *P* can be determined observationally, then, in optimal conditions, *#P#* may be tokened following an observation that (implies) *P*, and will be tokened *only* if *P*

(4) In optimal conditions, reflection on the contents of other beliefs which straightforwardly imply *P* will lead to a tokening of *#P#*

I am inclined to take (2) as a special case of (1), both of which can be taken as implications of the Belief-Desire Law. Following Stalnaker (1984), we can refer to (1) and (2) as *forward-looking* roles: they inform us as to the kinds of states that a *belief that P* typically brings about, whereas *backward-looking* roles (like (3) and (4)) inform us as to the kinds of states which typically bring about a *belief that P*.

The Belief-Desire Law is usually also taken to specify the characteristic functional role of *desiring that P* as well—that is, functional role semanticists will generally assert that *#Q#* underwrites a *desire that Q* if:

(5) In conjunction with a belief that *P*, *#Q#* leads to behaviour which would tend to bring it about that *Q* at worlds where *P* is true (along with all the subject's other beliefs)

Nothing like (3) and (4) seem to apply to *desires*, however—these two account for the special *epistemic* function that beliefs are supposed to play, whereas desires are usually taken to be characterised primarily in terms of their *motivational* function.

There are, in other words, two *basic* kinds of functions that beliefs are generally assumed to play, while desires perform just one: beliefs and desires jointly guide behaviour, and beliefs are also supposed to change in response to evidence and reasoning. The functionalist account of credences and utilities that I will suggest in the next section will have the same character, *mutatis mutandis*—though it is also compatible with credences and utilities playing other roles not yet mentioned. However, unlike a functional role semantics grounded in (1) to (5), which apply to *individual* beliefs and desires, I will provide functional roles in the first instance for *total* credence and *total* utility states.

In one form or another, (1) to (5) capture the most commonly cited roles associated with beliefs and desires. See, e.g., (Pettit 1993), (Lewis 1972, 1994), (Shoemaker 2003), and especially (Loar 1981), who emphasises versions of (1) and (3) in particular. Stalnaker also bases his account of belief on versions of (1) and (3) (and his account of desire on a version of (5)), asserting that:

> Very roughly, to believe that *P* is to be in a state that is sensitive to the information that *P*, and that disposes the agent to do what would best satisfy his desires if *P* (together with his other beliefs) were true. (1999a, 152)

It is worth pausing briefly on a small exegetical matter here, as it will help us to clarify the nature of functional role semantics. In his (1984), Stalnaker asserts that "Our [beliefs] represent what they represent not only because of the behaviour they tend to cause, but also because of the events and states that tend to cause them" (18); and later, that "Both the forward-looking and the backward-looking aspects of [beliefs and desires] are *essential to the explanation of how they can represent the world*" (19, emphasis added). These passages suggest a functional role semantics based primarily on (1) and (3), where *both* are treated as being important *vis-à-vis* the *content* of our attitudes.

At the same time, however, Stalnaker also sometimes seems to suggest that (1) and (3) have distinctive roles to play in the explanation of belief, with (3) fixing the *content* of the attitude and (1) fixing the *type* of attitude that it is (e.g., a belief rather than a desire):

> We believe that *P just because* we are in a state that, under optimal conditions, we are in only if *P*, and under optimal conditions, we are in that state because *P*, or because of something that entails *P*. But a causal account of belief … cannot, I think, *replace* [a pragmatic analysis of belief in terms of the Belief-Desire Law], it can only supplement it. For an account of belief must explain, *not only* how belief can represent the world, but also what distinguishes *belief* from other kinds of representation states … Beliefs have determinate content because of their presumed causal connections with the world. Beliefs are *beliefs* rather than some other representational state, because of their connection, through desire, with action. (1984, 18-19, emphasis added)

A natural reading of this passage is that Stalnaker proposes to explain the *content* of beliefs *via* indication relations, and then to distinguish beliefs from other representational states by means of their distinctive roles in cognition.[45] This is similar to the two-step strategy pursued by Boghossian, above, and distinct from the strategy that I want to pursue for characterising credences and utilities.

It is difficult to reconcile the foregoing passages, and I will not try to here. What I do want to note, however, is that a functional role semantics is *not* committed to supposing that any one role has explanatory priority with regards to content. In particular, the version of the view that I will suggest in the next section treats the *forward-looking* roles of credences and utilities (in particular, their connection with preferences) as being *semantically* important—even if, ultimately, backward-looking roles may be required as well.

Before we move on, I want to make two points about how functional role semantics should be cashed out. First of all, note that *if* functional role semantics is to going to supply an account of the *content* of an attitude ϕ, then the roles associated with ϕ must be capable of pinning down an appropriately *unique* content. For instance, while role (2) might *differentiate #P#* as a basis for a *belief* rather than a *desire*, we could not characterise a *belief that P* only using (2) as that role does not give us enough information to work out *#P#*'s content (or even very tightly constrain the possibilities). Likewise, most have argued that (1) and (5) *by themselves* aren't enough to functionally characterise what it is to *believe that P* and *desire that Q*, on the basis of informal arguments that suggest that any given pattern of behaviour is consistent with an extremely wide range of interpretations consistent with the Belief-Desire Law (§4.2). If a functional role semantics for ϕ is to get off the ground, then, a strong case needs to be made for thinking that the roles associated with ϕ can fix upon a unique assignment of contents (and, of course, that they fix upon the *right* contents).

Secondly, functional role semantics (as I am here understanding it) need not restrict itself to straightforwardly causal roles—that is, roles of the form "*ϕ-ing that P* causes *x*" and "*ϕ-ing that P* is caused by *y*". The examples (1) to (5) are more plausibly understood as *normative* roles, in one sense of 'normative' or another. There are two obvious ways to naturalistically cash out this notion of 'normativity'. The first is *statistical*: (1) to (5) characterise statistically normal causal connections associated with the psychologically real entities which underlie our beliefs and desires; e.g., a *belief that P typically* (but not always) leads to behaviour which tends towards desire satisfaction at worlds where *P* is true. Alternatively, one can appeal to the role that a *belief that P* (typically) plays in a *typical* member of the population/species (cf. §3.3.1).

[45] Interestingly, *desires* are said to "have determinate content because of their dual connection with belief and action" (19)—not because of any prior causal connections or other backward-looking connections.

The second kind of normativity is *biological*: it is not implausible that (1) to (5) characterise the *proper functional roles* of beliefs and desires; e.g., a *belief that P* would, were it functioning properly under ancestrally normal conditions, lead to behaviour which tends towards desire satisfaction at worlds where *P* is true—regardless of whether the belief brings that kind of behaviour about *as a matter of fact*. One might, therefore, combine functional role semantics with a kind of non-compositionalist teleosemantics, if the ‘roles’ involved in characterising the attitude are cast in terms of those state’s proper functions. For an example of this strategy, see (Papineau 1984, 1987), where appeal is made to the proper functions of whole belief and desire states in order to account for the contents of those states in a non-compositional manner.

4.5 Realist characterisational representationism

It is towards the development of a functional role semantics for credences and utilities that a representation theorem *of the right kind* would seem especially useful for the psychological realist. Any such theorem will:

(i) Have preference conditions that are at least approximately satisfied by the majority of (properly functioning) ordinary agents
(ii) Have a reasonably strong uniqueness condition
(iii) Establish a representation scheme with complete models of agents’ credences and utilities which fit reasonably well with the intuitive and empirical data

For what follows, it will be helpful to keep in mind the distinction between *mentalistic* and *behavioural* preferences (§2.2). Representation theorems (or more accurately, their Decision-theoretic Interpretations) can be distinguished by the kind of preferences to which their conditions refer—there are those which are built around a behavioural interpretation of $\succcurlyeq$, and those built around a mentalistic interpretation. I will discuss the consequences of this distinction in more detail below; for now, I will simply speak in terms of ‘preferences’ without specifying the kind.

The basic idea behind the kind of psychologically realist characterisational representationism that I have in mind is functionalist, where credences and utilities are identified at least in large part through their explanatory role in the production of intentional behaviour. Representation theorems could be used to precisely spell out either the content-determining functional role associated with our credence and utility states, or at least a very important *part* of that role. In particular, a representation theorem of the right kind could be used to supply a joint role for an agent’s total credence and total utility states, which—depending on the strength of the theorem’s uniqueness condition—is capable of either fully determining the content for those states, or at least narrowing that content down to a relatively small range of possibilities.

Consider, for example, a theorem *T*—let us suppose it's an otherwise standard expected utility theorem with a non-probabilistic credence function—which satisfies (i) to (iii) with the Standard Uniqueness Condition. Nothing hinges on whether *T* is an expected utility theorem, but supposing as much will make the following discussion more straightforward. The example can easily be modified for NCU theorems—as noted in §3.3, almost every representation theorem which has been developed in the last 100 years leaves us either with an expected utility model, or something which comes very close to $\mathcal{EU}$-maximisation—e.g., $\mathcal{EU}$-maximisation with some fudge-factor that accounts for risk aversion. By hypothesis, *T* allows us to pair the total preference patterns of ordinary agents with what is an effectively unique representation of that agent as an expected utility maximiser, where that representation corresponds closely to our intuitions regarding what credences and utilities the agent in question might actually have under those conditions.

To the extent that we are psychological realists, and thus think that ordinary agents' credences and utilities are underwritten by *M*-representations, establishing *T* paves the way for a functional role semantics based on the following joint roles:

(6) A total credence state $\mathcal{B}el$, in combination with a total utility state $\mathcal{D}es$, (typically) leads to a preference system $<\boldsymbol{\mathcal{BOP}}, \succcurlyeq>$ such that $x \succcurlyeq y$ iff the expected utility of *x* is greater than the expected utility of *y*

(7) A total preference system $<\boldsymbol{\mathcal{BOP}}, \succcurlyeq>$ which satisfies *T*'s preference conditions is (typically) caused by a total credence state $\mathcal{B}el$ in combination with a total utility state $\mathcal{D}es$.[46]

We don't need the representation theorem to establish (6); simple mathematics is enough to establish that, given $\mathcal{B}el$, $\mathcal{D}es$, and the posited $\mathcal{EU}$-maximisation rule, $<\boldsymbol{\mathcal{BOP}}, \succcurlyeq>$ will have such-and-such a structure. However, the theorem does allow us to establish (7), which is needed to ensure that the specified functional roles are capable of pinning down an effectively unique assignment of credences and utilities. Contrast this with the use of the Belief-Desire Law to functionally characterise beliefs and desires, where the standard complaint is that these roles are compatible with far too many interpretations.

(6) and (7) specify roles for *total* credence and utility states, but are neutral with respect to how those states must be psychologically realised. They are therefore consistent with a map-like view on the structure of thought, where each agent's total credence and utility states are underwritten by just one, or relatively few, informationally rich *M*-representations. They are also consistent with a sentential view, in the sense that they suggest that contents can be assigned to whole collections of sentences at once. If the sentential view were correct, then, the task would be to use that assignment to determine the contents of

[46] (6) and (7) are not equivalent: $\mathcal{B}el$ and $\mathcal{D}es$ could typically bring about a particular system of preferences $<\boldsymbol{\mathcal{BOP}}, \succcurlyeq>$, without it being the case that $<\boldsymbol{\mathcal{BOP}}, \succcurlyeq>$ is typically the result of $\mathcal{B}el$ and $\mathcal{D}es$ (e.g., if there are many other things which often lead to $<\boldsymbol{\mathcal{BOP}}, \succcurlyeq>$).

individual sentences (and concepts), presumably by considering the role that each such sentence plays with respect to the whole and working backwards from there.

As was noted with (1) to (5) above, (6) and (7) may be construed as *normative* roles, rather than straightforward causal roles. For instance, it's not implausible to suppose that something like (6) specifies the causal role of our total credence and utility states when they are *functioning properly* in normal conditions—in which case we would not expect that having preferences $<\mathcal{BOP}, \succcurlyeq>$ which satisfy *T*'s preference conditions would *automatically* qualify an agent as having credences $\mathcal{B}el$ and utilities $\mathcal{D}es$ (as would be implied under CCR). Current scientific models of decision-making tend to idealise away from factors known to interfere with our deliberative capacities, such as intoxication and so on; and they are only intended to model *typical* subjects. This weakens the posited connection between preferences, credences, and utilities to one that only holds under the right conditions, but that should not stop us from using such weakened connections to characterise credences and utilities.

Note, also, that if *T*'s uniqueness condition were not as strong as the Standard Uniqueness Condition, then further functional roles would have to be called upon to pin down appropriate contents. Indeed, *even if* (6) and (7) managed to pin down a unique model of the agent's credences and utilities, I expect that it would be valuable to take into account other roles besides—after all, neither (6) and (7) take into account how an agent's credences do (or should) change in response to evidence and reasoning, which seems to be around about as important for the understanding the nature of credences as is their role in the production of preferences (§3.3.1).

An analogy may be helpful here. Suppose the task is to outline the meaning of the term 'water', which I will assume can be best accounted for by some form of causal descriptivism.[47] (I have defended descriptivism elsewhere; see Elliott, McQueen *et al*. 2013.) We might begin with the description *D*, that water is *the potable, clear liquid around here which comes out of our taps that we need to drink to survive*. That would probably be enough to fix the referent in this world and in most of the nearby possibilities that we might consider—but it doesn't tell us *everything* there is to the meaning of 'water'. There are many other properties associated with our use of the term which aren't mentioned in that short description; e.g., *fills the lakes and oceans*, *falls from the sky as rain*, *boils at 100° C* and *freezes at 0° C*. There is more packed in to our concept than we need to pin down the referent. Roughly speaking, *D* captures a large and centrally important *chunk* of the meaning of 'water', but it leaves a lot out as well; and there is no particular reason

[47] That is, the meaning of 'water'—or at least one of its meanings—can be given by a (potentially infinite) description which uniquely identifies *water* across a range of possible scenarios considered as actual (Lewis 1984, 1994, Kroon 1987, Jackson 1998). The description is generated *via* a collection of properties (or sometimes: platitudes) that the speaker associates with their use of the term.

to think that *D* should be taken as paramount when other descriptions could also do the job.

Likewise, appeal to other functional roles would not imply the insignificance of (6) and (7) for the characterisation of credences and utilities. Credences have an epistemic role to play as well, besides their role in guiding behaviour, but *both* kinds of role are central to understanding what credences *are*. It is, I think, too much to expect any one representation theorem that it provide the *whole* story about what it is to have such-and-such credences and utilities—especially in light of the fact that, as emphasised in §3.3.1, having credences and utilities is *not* simply a matter of having particular preference patterns. But where credences and utilities are to be understood and characterised in terms of the *roles* that they play (or are supposed to play), something like (6) and (7) very plausibly form an important *part* of what it is to have those attitudes.

Let us close by considering how this realist position relates to the naturalisation project. As I have explained it, a functional role semantics is not committed to naturalisability—although, as a matter of fact, most functional role semanticists have adopted the position in their search for a fully naturalistic account of the attitudes. In the long run, the naturalistic functional role semanticist will want to cast everything in terms of external causal inputs and behavioural outputs, with all reference to intentional or otherwise mental phenomena having been Ramseyfied away (see Lewis 1970, 1972).[48]

To this end, however, a naturalistic functional role semantics for *beliefs* and *desires* is but a twinkle in the eyes of some philosophers. It is clear that the commonly noted functional roles for belief and desire are not cast in naturalistic terms: each of (1) to (5) refer to other intentional states, and it's plausible that reference would need to be made to other mental states to spell out the 'optimal conditions' mentioned in (3) and (4). As is widely recognised, the causal properties of any one mental state will usually depend on the presence or absence of a range of other mental states, and no Ramseyfication can exist without a complete specification of the relevant causal role of each of the many interconnected mental states which interact to determine any one mental state's causal properties. We are a long way from giving any such specification for beliefs and desires. At best we have just a rough idea of how our beliefs and desires connect to behaviour and to the non-intentional world more generally.

[48] Not all philosophers who pursue a broadly functionalist approach wish to further the naturalisation project. Schwitzgebel (2002, 2013), for instance, explicitly opts to set aside naturalisation, and argues instead for what he calls *liberal dispositionalism* (see also Baker 1995). In outline, Schwitzgebel's view is that beliefs *are* dispositions (or collections of dispositions), including dispositions to act in such a way as to tend to bring about what one desires *à la* the Belief-Desire Law. However, liberal dispositionalists allow for the characterisation of what it is for *S* to *believe that P* to be given partly in terms of other propositional attitudes and mental states—including, potentially, other beliefs—while making no promises to eventually naturalise away any reference to those states.

Part of the problem here is that a functional role semantics for belief and desires has no way of working backwards from facts about behaviour to facts about our beliefs and desires. Appeal to the Belief-Desire Law just doesn't allow us to constrain the possible assignments of beliefs and desires tightly enough. The promise of a solution to this problem accounts for much of the appeal of characterisational representationism. Indeed, I suspect a great deal of progress could be made towards a naturalistic, functionalist construal of credences and utilities *if* we could prove a representation theorem of the right kind, which took us from a typical subject's *behavioural preference system*—characterised in purely naturalistic terms—to a unique and plausible assignment of credences and utilities. The development of such a theorem would at least allow us to take steps towards a completely naturalistic reduction of credences and utilities.

Unfortunately, as will become clear in the chapters that follow, such a theorem has yet to be developed. Indeed, it does not appear that decision theorists have even come very *close* to developing a theorem appropriate for such purposes. We may, *one day*, have a representation theorem that is well-suited for advancing the naturalisation project, but it will probably not look much like any of the theorems which exist today. In particular, it will probably not involve preferences over *act-functions* or *lotteries*, for reasons to be discussed in Chapter 5 and Chapter 6. Act-functions and lotteries form the standard way of characterising the basic objects of preference in any representation theorem geared towards a behavioural interpretation of $\succcurlyeq$, but they also lead to the most worrying issues with those theorems with respect to their application to characterisational representationism. The more plausible option, *given the theorems we currently have*, would be to appeal to a theorem which specifies conditions on mentalistic preferences (see §6.2 and §8.3)—or hold out hope for a new and better theorem. Either option, however, means putting the naturalisation project on hold, at least for a time.

Of course, to develop a functional role semantics with (6) and (7) characterised in terms of mentalistic preferences is not *incompatible* with the naturalisation project—it merely fails to clearly advance that project. (Compare: characterising beliefs in terms of (1) to (4) does not entail that they cannot be naturalised, but neither does it immediately point the way to naturalisation.) Perhaps the naturalistic philosopher could seek to characterise credences and utilities in terms of mentalistic preferences, offering a promissory note to naturalise mentalistic preferences at some point down the line—after all, some such promissory note has been offered by every purportedly naturalistic account of beliefs and desires yet developed.

I will have more to say on whether the naturalisation of mentalistic preferences is feasible in Chapter 9; in the interim, the question is whether *any* current representation theorem has the right properties to be a plausible foundation for characterisational representationism.

4.6 Summary

There were three main lessons drawn in Chapter 3. First, credences and utilities are not just preference states, nor does it appear that having any particular pattern of preferences is sufficient for having such-and-such credences and utilities. Credences in particular play an epistemic role, and an adequate account of what they are should accommodate this fact. Secondly, the proponent of characterisational representationism ought to avoid theorems with preference conditions that ordinary agents do not come close to satisfying. And thirdly, she also ought to avoid theorems with excessively restrictive representational resources.

An appeal to a theorem with the right properties would ensure that characterisational representationism stays in line with the final two of these lessons. And, as we have now seen, there are several ways to cash out characterisational representationism while keeping an agent's system of preferences metaphysically and conceptually distinct from her system of credences and utilities, while taking into account the special epistemic role that credences are supposed to play. Psychological non-realists aren't committed to Classical Characterisational Representationism, as they might (like Lewis and other interpretivists) appeal to information which goes beyond agents' preferences. The same, of course, can be said for psychologically realist versions of characterisational representationism, which might (a) appeal only to the normative (rather than actual) roles that credences and utilities have in the production of preference patterns, and/or (b) appeal also to other factors beyond agents' preferences. With the right representation theorem, characterisational representationism could avoid the main pitfalls that are notoriously associated with Naïve, Extreme, and Classical Characterisational Representationism.

Moreover, the foregoing review gives strong reason to take the characterisational representationist's approach seriously. We currently have no fully worked out account of beliefs and desires; instead, what we have is a number of rough ideas much in need of further development. A recurrent theme, though, is that we ought to be able to characterise the propositional attitudes by reference to what they *do* (or should do, or typically do, or do under certain conditions)—where one of the most important things that beliefs and desires do involves their role in the explanation of preferences and intentional action. It would be difficult to understate the importance of the Belief-Desire Law for most attempts to understand and characterise beliefs and desires. As we have seen, it is central to almost all varieties of Basic Psychological Realism regarding those attitudes, where it's used to characterise both belief-like and desire-like roles. It also forms a centrepiece for a functional role semantics for beliefs and desires, and for each of the two kinds of psychological non-realism that we looked at in §4.2.

When it comes to the metaphysics of credences and utilities, it seems fair to expect that the decision-theoretic analogue of the Belief-Desire Law—the principle of expected utility maximisation (or something very close to it)—is likely to play just as central a role.

While they might do other things besides, if credences and utilities do *anything*, they are closely connected to our preferences—and plausibly *via* something which looks roughly like expected utility maximisation. To have a theorem, then, which connects preference patterns to a very limited range of plausible credence and utility assignments, would seem a very useful resource for the precise functional characterisation of those attitudes.

CHAPTER FIVE

The Instability of Savage's Foundations

Savage's *The Foundations of Statistics* (1954) is centred around one of the most well-known and admired representation theorems ever developed. David Kreps describes Savage's theorem as the "crowning glory of choice theory" (1988, 120). Likewise, in summarising his widely-cited review of over two dozen CEU representation theorems, Peter Fishburn has this to say:

> Savage's [theorem] is suitable for a wide variety of situations, its axioms are elegant and intuitively sensible, and its representation-uniqueness result is easily connected to assessment techniques [. . .] I regard it as one of the best. (1981, 194)

The admiration for Savage's work shows through in its influence; indeed, it would not be unfair to characterise axiomatic decision theory since 1954 as a series of footnotes to Savage.[49] The majority of representation theorems—for both CEU and NCU—that exist today are based upon the same basic formal system as the one that Savage developed, usually with only minor tweaks here and there.

Despite all this—or perhaps because of it—Savage's *Foundations* has also attracted a lot of criticism. At the forefront of this critique is the so-called *constant acts problem*.[50] As we will see, it's not clear how much of a problem there is here, at least for characterisational representationism. Nevertheless, there are greater concerns on the horizon, which have their origins deep within the formal paradigm that Savage developed and affect every theorem based on his system.

It would be impossible to look at every representation theorem that falls within the Savage paradigm—these number well into the dozens. Instead, I will begin in §5.1 by describing Savage's formal framework and theorem in some detail. Following that, I will consider a number of reasons why Savage's theorem, and other theorems based on the same framework, are unsuitable as a basis for characterisational representationism. In

[49] Savage himself was greatly influenced by Bernoulli (1738), Ramsey (1931), de Finetti (1931, 1964), and von Neumann and Morgenstern (1944), amongst others.

[50] Two further complaints that are commonly made against Savage's theorem are that he requires his set of states to be uncountable, and the so-called *problem of small worlds*, neither of which I will discuss here. See (Joyce 1999, 70-7, 110-13) for a thorough discussion of the latter.

§5.2 I focus on the constant acts problem, whereas in §5.3 and §5.4 I consider what I take to be two more fundamental issues with Savage's framework.

5.1 Savage's *Foundations*

My exposition of Savage's theorem will be in two parts. In §5.1.1, I begin with a relatively informal characterisation of the basic elements needed to understand his theorem, and then in §5.1.2, I outline Savage's preference conditions and say a few words about the final representation result.

5.1.1 Preliminaries

According to Savage, the basic objects of preference are *acts*. Intuitively, acts are the kinds of things that we might choose to do in a given decision situation. For instance, when bored, one might choose to *read a book* or *go fishing*; at night, one might *go to bed* or *stay awake*; in a game of poker, *hold* '*em* or *fold* '*em*. We cannot choose, however, to *slow the speed of light*, nor *stop the Earth spinning*: such things we could not realise even if we intended to, so in an intuitive sense they are not acts available to us. It's difficult, however, to go very far beyond this rather vague gloss on what acts are exactly, and I will not try to here. For now, I will adopt the intuitive notion of an act, though I will have more to say on the issue later.

Suppose we have a non-empty set, $\mathcal{A}'$ = {α, β, γ, …}, containing a range of acts available to some subject *S* in an unspecified decision situation. As only one act in $\mathcal{A}'$ can ever be realised by the decision-maker, $\mathcal{A}'$ should be understood as containing act types rather than tokens. Alternatively, one could think of $\mathcal{A}'$ as a set of propositions which specify that *S* performs one of the acts available to her—there are no important issues that arise from construing $\mathcal{A}'$ as a set of acts or propositions about acts.

There are three things that need to be said about how $\mathcal{A}'$ is to be specified. First of all, every act α in $\mathcal{A}'$ should be such that *S* is certain that she would perform α, if she were to intend as such. For instance, *S* might intend to *travel to New York*, but whether she succeeds or not depends on a number of factors outside of her control which could, for all she knows, prevent her from arriving. On the other hand, in most cases she can, say, *reach for the nearest object*, and she can be sure that she will succeed in doing so should she so choose. Secondly, acts can be described at different levels of specificity; for instance, to *read Moby Dick* is one way to *read a book*, but it's not the only way. I will assume that the acts in $\mathcal{A}'$ are specified at least at a reasonably fine-grained level. And finally, $\mathcal{A}'$ should be specified in such a way that *S* must perform at least one act in $\mathcal{A}'$, and the performance of any one such act in $\mathcal{A}'$ should preclude the performance of any other. (Thus, if *read Moby Dick* is in $\mathcal{A}'$, *read a book* cannot be—but *read the Odyssey* might be.) The motivation for these restrictions on $\mathcal{A}'$ will be discussed in §5.4.

Savage's central motivation for characterising the basic objects of preference as acts was that he intended a behavioural interpretation for his use of $\succcurlyeq$. For Savage, an agent's preference ranking over acts is supposed to somehow directly encode her behavioural dispositions in choice situations, thus making her preferences—and hence her credences and utilities—open to empirical investigation (see, for example, Savage 1954, 27-30). As he put it, "Loosely speaking, [α] $\succcurlyeq$ [β] means that, if [the agent] were required to decide between α and β, no other acts being available, he would decide on α" (1954, 17).

Acts usually have a range of different *outcomes*, depending on the different *states* that the world might be in. If I read a book then I might either *become entertained* or *become annoyed*, depending on the (presently unknown to me) contents of its pages; and if I go fishing, I might *catch a fish* or *catch nothing*, depending on what's in the water. Let $\mathcal{O}$ = $\{o_1, o_2, o_3, \ldots\}$ contain descriptions of *each* of the possible outcomes that might arise given any act in $\mathcal{A}'$, focussed in particular on describing those states of affairs that *S cares about*. (I do not care, for instance, that if I *go fishing*, then I will still *have an even number of pencils in my office*, so we can leave that out of the description of the outcome.) As Savage describes the outcomes in $\mathcal{O}$, "They might in general involve money, life, state of health, approval of friends, well-being of others, the will of God, or anything at all about which the person could possibly be concerned" (Savage 1954, 14). For reasons to be clarified below, the descriptions ought to be fairly specific (if not maximally specific) with respect to what *S* cares about, and—importantly—they should be mutually exclusive. Since exactly one act in $\mathcal{A}'$ must be performed, the set of outcomes is jointly exhaustive of the possibilities.

Finally, we will need a set of the *states*, $\mathcal{S}$ = $\{s_1, s_2, s_3, \ldots\}$, upon which the different outcomes of *S*'s acts depend. The collection of states should be a *partition* of some possibility space (I will leave it open which space); i.e., a collection of propositions such that exactly one is true. Savage does not explicitly describe states in much detail. There are, however, two critically important properties that we need to assume states have if Savage's theorem is to have a plausible interpretation *qua* decision theory, which I will outline now.

First of all, states should be *independent* of whatever act the agent might choose to perform. In the literature, this property of states is referred to as *act-independence*. As Allan Gibbard and William Harper (1978) have pointed out, Savage's system is compatible with (at least) two notions of independence being applied in the precisification of this requirement. The first is *evidentially independence*, where a state *s* is *evidentially independent* of the performance of an act α just in case *S*'s credences that *s* is true under the assumption that she performs α is equal to her credences that s is true under the assumption that she does not perform α.[51] The second kind of independence they refer to as *causal*, though it would be better termed *counterfactual independence*. A state s is *counterfactual independent* of the performance of an act α just in case s would hold if α were performed, and s would hold if α were not performed.

For the purposes of the present exposition, it's not important which of these two notions of independence is used. I will, however, note a consequence of applying either—namely, that states must be *logically independent* of acts:

> **Definition 5.1: Logical independence**
> A state s is *logically independent* of the performance of an act α iff s is consistent with α being performed and α not being performed

This allows us to define the key property of *act-independence*:

> **Definition 5.2: Act-independence**
> A state s is *act-independent* (with respect to a choice of $\mathcal{A}'$) iff s is logically independent of the performance of any $\alpha \in \mathcal{A}'$

As Savage requires that every state in $\mathcal{S}$ is act-independent, a state cannot entail that a particular act in $\mathcal{A}'$ is chosen (or not chosen).

Secondly, states should be *outcome-functional*:

> **Definition 5.3: Outcome-functionality**
> A state s is *outcome-functional* (with respect to a choice of $\mathcal{A}'$ and $\mathcal{O}$) iff the performance of any s-compatible $\alpha \in \mathcal{A}'$ at s uniquely determines that a particular outcome $o \in \mathcal{O}$ obtains

The upshot of assuming outcome-functionality is that, for each state s, there will be a function which maps every act in $\mathcal{A}'$ which might be performed at s to an outcome in $\mathcal{O}$; if *every* act in $\mathcal{A}'$ is compatible with s, then it will be a total function on $\mathcal{A}'$. Note that

[51] Evidential independence is standardly characterised in terms of *probabilistic independence*; *viz.*, if $\mathcal{B}el$ is a probability function, then s is evidentially independent of the performance of α (relative to $\mathcal{B}el$) just in case $\mathcal{B}el(s|\textit{perform}\ \alpha) = \mathcal{B}el(s|\textit{don't perform}\ \alpha)$, where $\mathcal{B}el(P|Q) = \mathcal{B}el(P\ \&\ Q)/\mathcal{B}el(Q)$. If s is evidentially independent of all acts in $\mathcal{A}'$, which are by hypothesis mutually exclusive and jointly exhaustive, then for any act $\alpha \in \mathcal{A}'$, $\mathcal{B}el(s|\textit{perform}\ \alpha) = \mathcal{B}el(s|\textit{don't perform}\ \alpha) = \mathcal{B}el(s)$. I have avoided this formulation of evidential independence because of its use of conditional probabilities, the application of which raises concerns insofar as S isn't probabilistically coherent. There are some difficulties with the formulation of evidential independence given here, but the precise formulation is not important for the discussion that follows.

act-independence and outcome-functionality are not *formal* requirements on the specification of $\mathcal{S}$, which for the purposes of the theorem may be characterised sparsely as any non-trivial partition of non-empty set. Rather, act-independence and outcome-functionality are two properties that we must assume the states in $\mathcal{S}$ have, if Savage's theorem is to have a plausible interpretation *qua* decision-theory.

In Savage's framework, states are the ultimate objects of uncertainty: it is from $\mathcal{S}$ that Savage constructs the domain of his $\mathcal{B}el$ function—namely, the set of *events*, $\mathcal{E} = \{E_1, E_2, E_3, \ldots\}$. Each event is a set of states, and Savage assumes that every set of states is included in $\mathcal{E}$ (i.e., $\mathcal{E} = 2^{\mathcal{S}}$). Although events are technically sets of states rather than propositions *per se*, we do no harm in treating events as propositions. As all states are pairwise inconsistent, every event corresponds directly to one and only one proposition, *viz*., the disjunction of each of the states in the event. We will therefore treat events as propositions. It should be clear, given this characterisation of events, that they inherit the event-equivalent property of act-independence from the states of which they are composed (but they don't inherent anything like outcome-functionality).

Savage's central insight was the recognition that, given the way we have characterised $\mathcal{S}$ and $\mathcal{O}$, each act in $\mathcal{A}'$ can be uniquely modelled by a function form $\mathcal{S}$ to $\mathcal{O}$. The idea is that each such function determines a unique definite description that identifies a particular act that the agent might perform—or at least a class of acts which are, from the perspective of the decision-maker, not worth distinguishing:

> If two different acts had the same consequences in every state of the world, there would from the present point of view be no point in considering them two different acts at all. An act may therefore be identified with its possible consequences [at different states of the world]. (1954, 14)

(Of course, if the outcomes are specified in enough detail, it's highly unlikely that two acts would have the same outcomes across all states.) Suppose that $\mathcal{F}$ is the function that pairs the state s_1 with the outcome o_1, s_2 with o_2, and so on; we can then say that $\mathcal{F}$ represents:

> the act α in $\mathcal{A}'$ such that, were it performed, then (if s_1 were the case, o_1 would result) & (if s_2 were the case, o_2 would result) & …

We will refer to any function from a set of states to outcomes as an *act-function*. Savage's $\succcurlyeq$ is formally defined on a set of act-functions, and it's this feature which essentially characterises the influential formal paradigm he developed. For most theorems within this paradigm, act-functions are total functions on $\mathcal{S}$ and often only take a finite number of values from $\mathcal{O}$. In the literature, act-functions are often called *Savage acts*; however it will be helpful for the discussion that follows to distinguish the *functions* and the *acts* that they supposedly represent.

Note that the representation of acts as *total* functions from $\mathcal{S}$ to $\mathcal{O}$ would be nonsensical if some states were logically incompatible with the performance of some acts—what sense would it make to speak of an act's outcome at a state which *implies* that the act is not performed? Likewise, outcome-functionality is required if a function from states to outcomes is to represent an act along the lines described—if, for example, α could only ever result in either o_1 or o_2, but every state in $\mathcal{S}$ left it indeterminate which of these outcomes would result, then there would be no reason to suppose that α corresponds to one function from $\mathcal{S}$ to $\{o_1, o_2\}$ rather than any other.

With the set of events specified as the set of all subsets of $\mathcal{S}$, it's worth noting that every one of Savage's act-functions can be expressed equivalently as a mapping from a set of mutually exclusive and jointly exhaustive *events* to outcomes, simply by collecting together the states with similar outcomes into a single event. For example, if $\mathcal{F}(s) = o_1$ for all states s in E, and $\mathcal{F}(s) = o_2$ for all states s in $\neg E$, then we might represent $\mathcal{F}$ as $(E, o_1 | \neg E, o_2)$. More generally, assume the following convention for representing act-functions:

Definition 5.4: Act-function notation
$\mathcal{F} = (E_i, o_i | \ldots | E_n, o_n)$ iff $\{E_i, \ldots, E_n\}$ is a partition of $\mathcal{S}$ and if $s \in E_i$, $\mathcal{F}(s) = o_i$, …, and if $s \in E_n$, $\mathcal{F}(s) = o_n$

This convention will be helpful in laying out Savage's preference conditions and formal results more transparently.

So far, I have treated *acts* as a kind of conceptual primitive, with *states*, *outcomes*, and *events* being partially characterised by their relation to the acts in $\mathcal{A}'$. In Savage's formal system, however, the situation appears rather different. Savage begins with two primitive sets: $\mathcal{O}$ and $\mathcal{S}$. Formally, all that is required of $\mathcal{O}$ is that it contains at least two members, of $\mathcal{S}$ that it is a non-trivial partition of some non-empty set—sparse characterisations, to be sure, but this hides the informal properties they must have if they are to stand for collections of outcomes and states respectively. There is no *formal primitive* which corresponds to $\mathcal{A}'$. Rather, from $\mathcal{S}$ and $\mathcal{O}$, Savage constructs the set which we will label $\mathcal{A} = \{\mathcal{F}, \mathcal{G}, \mathcal{H}, \ldots\}$, which contains all total functions from $\mathcal{S}$ to $\mathcal{O}$ (i.e., $\mathcal{A} = \mathcal{O}^{\mathcal{S}}$).

The construction of $\mathcal{A}$ from $\mathcal{S}$ and $\mathcal{O}$ is perhaps the most influential part of Savage's formal system (and, as we will see, the origin of its biggest problems). However, the order of the construction is somewhat misleading—suggesting as it does that acts can be straightforwardly *defined* in terms of states and outcomes. This is not at all the case, as the informal characterisations of $\mathcal{S}$ and $\mathcal{O}$ above highlight. Outcomes are characterised as the possible consequences of performing an act in $\mathcal{A}'$ under different states of the world, and states are characterised as necessarily consistent with the performance of any act in $\mathcal{A}'$ and such that the performance of any act in $\mathcal{A}'$ determines a unique outcome. There is no sense to be made of $\mathcal{S}$ and $\mathcal{O}$ as sets of states and outcomes *as they were described above* without a specification of $\mathcal{A}'$. There is, therefore, a sense in which the set of acts proper, $\mathcal{A}'$, is a kind of *informal primitive* which underlies any Decision-theoretic Interpretation of Savage's formal system.

Given that the states in $\mathcal{S}$ are act-independent and outcome-functional (with respect to a choice of $\mathcal{A}'$ and $\mathcal{O}$), it's clear that every act in $\mathcal{A}'$ can be uniquely represented by a particular act-function in the manner described above. It's far less clear, however, that every possible act-function in $\mathcal{A}$ corresponds a member of $\mathcal{A}'$. Nevertheless, Savage assumes that *all* act-functions are in $\mathcal{A}$—including, famously, *constant act-functions*. That is, for every outcome o in $\mathcal{O}$, there is a constant act-function in $\mathcal{A}$ that maps every state in $\mathcal{S}$ to o. It will be helpful to have special notation for constant act-functions:

Definition 5.5: Constant act-functions
$\underline{o} = \mathcal{F}$ iff $\mathcal{F}(s) = o$ for all $s \in \mathcal{S}$

Assuming that the outcomes are specified rather finely—as they must be, for reasons we will return to shortly—it's extremely doubtful that any constant act-function could serve to represent anything *real* that an agent might choose to do: what acts are there which would bring about any given outcome, regardless of how the world turns out to be? Nothing in the pre-theoretic, intuitive construal of the space of possible acts seems to have this character. In a nutshell, this is the *problem of constant acts*, which I will discuss in §5.2.

Constant act-functions play a number of important roles in Savage's theorem. For instance, Savage uses preferences between constant act-functions to construct a relative utility ranking upon the set of outcomes, which eventually gives rise to the utility function $\mathcal{D}es$—the idea being that the subject prefers the constant act $\underline{o}_1$ to $\underline{o}_2$ just in case she attaches a higher utility to o_1 than to o_2. This idea finds application then in Savage's definition of a relative credence relation, $\succcurlyeq^b$, defined on the space of events. The construction of $\succcurlyeq^b$ from $\succcurlyeq$ is crucial for the existence of Savage's $\mathcal{B}el$ function. In the literature, this has come to be known as Savage's principle of *Coherence*:

Definition 5.6: Coherence
For all $E_1, E_2 \in \mathcal{E}$, $E_1 \succcurlyeq^b E_2$ iff, for any $o_1, o_2 \in \mathcal{O}$, if $\underline{o}_1 \succcurlyeq \underline{o}_2$ then $(E_1, o_1 \,|\, \neg E_1, o_2) \succcurlyeq (E_2, o_1 \,|\, \neg E_2, o_2)$

This highly influential principle is *prima facie* intuitive—at least on the assumption that $(E_1, o_1 \,|\, \neg E_1, o_2)$ and $(E_2, o_1 \,|\, \neg E_2, o_2)$ actually correspond to things the agent can do. Suppose that the agent finds o_1 more desirable than o_2. So, if she is given a choice between two acts which each might result in either o_1 or o_2 but under different circumstances, our subject should prefer the act which, from her perspective, has the greater likelihood of resulting in o_1, and the smaller likelihood of resulting o_2. If she finds E_1 more likely than E_2 then, accordingly, she should find $(E_1, o_1 | \neg E_1, o_2)$ to be the more desirable act than $(E_2, o_1 | \neg E_2, o_2)$.

Of course, the foregoing reasoning requires the presupposition that o_1 obtaining under any state in E_1 is exactly as valuable for the subject as o_1 obtaining under any state in E_2, and likewise for o_2 in $\neg E_1$ and o_2 in $\neg E_2$. However, suppose that the following scenario occurs:

(a) S considers E_1 to be exactly as likely as E_2, i.e., $E_1 \sim^{b} E_2$
(b) S prefers the constant act $\underline{o}_1$ to the constant act $\underline{o}_2$
(c) S is generally indifferent between o_2 given $\neg E_1$ and o_2 given $\neg E_2$
(d) S finds o_1 substantially more desirable on average if it obtains in one of the states in E_1 than if it obtains in one of the states in E_2

Such a situation seems coherent; yet, in this case, presumably, the rational choice for S would be to prefer $(E_1, o_1 | \neg E_1, o_2)$ to $(E_2, o_1 | \neg E_2, o_2)$, despite the fact that $E_1 \sim^{b} E_2$. Although both acts have an equal subjective likelihood of resulting in o_1 and o_2, for the former act the outcome o_1 is much more desirable to S because it obtains in the right kinds of states. If o_1 can have a different subjective value for the agent if it obtains in any of the states in E_1 than it does if it obtains in any of the states in E_2, and similarly for o_2, then the justification for Coherence falls apart.[52]

Thus, it is frequently noted in the literature that Savage's theorem requires that outcomes are *state neutral*, where an outcome o is *state neutral* (relative to an agent S and specification of states $\mathcal{S}$) just in case S's utility for o does not depend on the state $s \in \mathcal{S}$ in which it's realised. However, simply requiring state neutrality is not *quite* enough to fully justify Coherence, which requires that the choice between $(E_1, o_1 | \neg E_1, o_2)$ and $(E_2, o_1 | \neg E_2, o_2)$ depends *solely* on the (presumed constant) values for o_1 and o_2, and the relative likelihoods of E_1 and E_2. To begin with, note that state neutrality does not yet rule out that the utility of an outcome may depend upon the *act* which gave rise to it. Thus, something stronger than state neutrality is needed, which I will call *context neutrality*:

Definition 5.7: Context neutrality
An outcome o is *context neutral* (relative to an agent S and a choice of $\mathcal{S}$ and $\mathcal{A}'$) iff S's utility for o depends neither on the state $s \in \mathcal{S}$ in which it's realised nor on the act $\alpha \in \mathcal{A}'$ from which it originates

[52] The same can be said for the definition of null events, and for the conditions **SAV3**, **SAV4** and **SAV5**, all discussed below.

Even the assumption of context neutrality is not quite enough, though, for it's conceivable that acts *themselves* could be objects of utility independently of their potential consequences. Thus Savage is forced to make an assumption about how agents value *acts*; namely, that they have no *intrinsic* preferences between acts, or preferences which don't depend upon the possible outcomes that the act might have. Without this assumption, it could be the case that the subject prefers o_1 to o_2, finds E_1 more likely than E_2, yet has such a strong *intrinsic* distaste for the act represented by $(E_1, o_1 | \neg E_1, o_2)$ that she is disposed to prefer $(E_2, o_1 | \neg E_2, o_2)$ instead *despite* its having the smaller likelihood of resulting in the best outcome.

Without these two assumptions, Savage's system becomes highly implausible, both descriptively and normatively. A natural thought here is that if agents care about the specific acts they perform, then that such-and-such an act was performed can be built into the description of the outcomes that obtain. Indeed, the most straightforward way to ensure the aforementioned requirements hold is to treat outcomes as conjunctions of states and acts. If outcomes are characterised in this way, then context neutrality is ensured and we don't need to assume that agents have no intrinsic preferences for acts.

However, this move does not sit well with other aspects of Savage's system (Joyce 1999, 56). Note, first of all, that since every outcome gets paired with every state by at least one act-function, and assuming that every act-function represents an act in $\mathcal{A}'$, it follows that states must be *outcome-independent* in the following sense:[53]

Definition 5.8: Outcome-independence
A state s is *outcome-independent* (with respect to a specification of outcomes, $\mathcal{O}$) iff s is logically consistent with any outcome $o \in \mathcal{O}$

For example, an outcome o cannot imply that a particular state s does *not* obtain, since (it is assumed that) there is some act the agent could perform which would bring about o if s were to be the case. Secondly, since every outcome is in the range of multiple act-functions, no outcome can imply that a *particular* act was chosen (though every outcome will imply that some range of acts was *not* chosen).

Thus, if the descriptions in $\mathcal{O}$ are intended to specify the various things the decision-maker may care about, the implication here is that the decision-maker has no intrinsic interest in what act she performs. This is, of course, also in the background of Savage's assertion that two acts with the same outcomes at all states are not worthy of being distinguished. Roughly put, Savage assumes that, from the decision-maker's perspective, *only potential outcomes matter*: the final decision model is one where the choice between acts depends wholly upon the credence-weighted utility of the outcomes; utilities for

[53] As with act-independence, events will inherit their own form of outcome-independence from states.

states and for acts *themselves* don't figure in the representation, which has a utility function defined only for the very limited set of propositions $\mathcal{O}$.

A number of authors have objected to the assumption of state neutrality (and by extension, context neutrality). (See, for instance, Karni, Schmeidler *et al.* 1983, Schervish, Seidenfeld *et al.* 1990, Bradley 2001.) I will not go over those complaints here; though I will note that *if* context neutrality is to be considered problematic, this can only be because it's in tension with other parts of Savage's system—context neutrality itself seems hardly problematic. Context neutrality forces outcomes to be rather fine-grained, and it's because of this that the problem of constant acts exists (see also the discussion in §5.2.1). To use an example of James Dreier's,

> I would rather have money as a gift from Boris than money stolen from Boris. The two outcomes must be distinguished. No one could plausibly accuse me of having intransitive preferences on the grounds that I preferred \$100 as a gift from Boris to \$5 as a gift from Boris, and \$5 as a gift from Boris to \$100 stolen from Boris. (1996, 257)

Here, Dreier is highlighting the distinction between characterising outcomes in a coarse-grained way,

o = *obtain \$100 from Boris*

And characterising them in a relatively fine-grained way,

o_1 = *obtain \$100 as a gift from Boris*
o_2 = *obtain \$100 stolen from Boris*

Most would value o_1 over o_2. However, an act whose outcome could be coarsely described as simply o may actually have outcomes manifest in particular as either o_1 or o_2, depending on the state of the world in which it's performed. Likewise, two distinct acts which both result in o given at a particular state may, more specifically, result in o_1 on the one hand or o_2 on the other. As the example highlights, the coarse-grained description of outcomes does not sit well with the presumption of context neutrality: the value of an outcome depends on the context in which it obtains; the more that context is built into the outcome, the less its value depends on outside factors. The idea here obviously extends beyond this rather simple example, suggesting that context neutrality is plausible only insofar as the outcomes in $\mathcal{O}$ are specified in rather great detail. Of course, given outcome-functionality, context neutrality then implies that $\mathcal{S}$ must be correspondingly fine-grained.

The following summarises the essential points to keep in mind for the critique which follows:

(1) $\mathcal{A}'$ = {α, β, γ, …} is a set of mutually exclusive *acts*, including all of the acts available to the agent in her present decision situation. Every act should be such that the decision-maker is certain that she would perform the act, if she were to so choose. It's assumed that agents have no intrinsic preferences between acts.

(2) $\mathcal{O} = \{o_1, o_2, o_3, \ldots\}$ is a set of *outcomes*; that is, a set of mutually exclusive and jointly exhaustive propositions about the consequences of performing an act at a state. For Savage's system to have a plausible interpretation *qua* decision theory, then the outcomes in $\mathcal{O}$ must be context-neutral and thus fine-grained, and they cannot imply that a particular act was chosen or that a particular state obtains.

(3) $\mathcal{S} = \{s_1, s_2, s_3, \ldots\}$ is a set of *states*; that is, a set of mutually exclusive and jointly exhaustive propositions. For Savage's system to have a plausible interpretation *qua* decision theory, the states in $\mathcal{S}$ must be act-independent in either the causal or evidential sense, and therefore logically independent of what acts are performed; they must also be outcome-functional. Together with the assumption that $\mathcal{A} = \mathcal{O}^{\mathcal{S}}$, the foregoing implies that states are outcome-independent.

(4) $\mathcal{E} = \{E_1, E_2, E_3, \ldots\}$ is a set of *events*; that is, (effectively) a set of propositions equivalent to disjunctions of states. Events inherit act-independence and outcome-independence properties from states.

(5) $\mathcal{A} = \{\mathcal{F}, \mathcal{G}, \mathcal{H}, \ldots\}$ is the set of all *act-functions*; that is, the set of all functions from $\mathcal{S}$ to $\mathcal{O}$. Such functions are intended to represent acts in $\mathcal{A}'$, by specifying the act's outcomes under different states.

(6) $\succcurlyeq$ is primitively defined on $\mathcal{A}$, and given a choice-based behavioural interpretation.

5.1.2 Savage's theorem

With all this in mind, we can now outline Savage's theorem and the structure of its proof. The theorem has seven preference conditions in the original formulation, though I will follow Joyce (1999) in explicitly listing the purely structural assumption that Savage needs to make about $\mathcal{A}$:

SAV0 $\mathcal{A} = \mathcal{O}^{\mathcal{S}}$

It's possible to weaken **SAV0** (and drop Savage's seventh preference axiom, **SAV7**) if we only desire the representation to hold for finitely-valued act-functions. In what follows, let $\mathcal{F}_E$ refer to the restriction of $\mathcal{F}$ to E. (Thus $\underline{o}_E$ is the restriction of $\underline{o}$ to E.) Furthermore, the *mixture* of $\mathcal{F}$ and $\mathcal{G}$, $\mathcal{F}_E \cup \mathcal{G}_{\neg E}$, is an act-function $\mathcal{H}$ such that $\mathcal{H}(s) = \mathcal{F}(s)$ for all $s \in E$, and $\mathcal{H}(s) = \mathcal{G}(s)$ for all $s \notin E$. We can now state the weakened act-richness assumption as follows:

SAV0’ $\mathcal{A}$ is the set of all finite-valued functions from $\mathcal{S}$ to $\mathcal{O}$; i.e., for any outcome $o \in \mathcal{O}$, $\underline{o} \in \mathcal{A}$, and for all $\mathcal{F}, \mathcal{G} \in \mathcal{A}$, and any $E \in \mathcal{E}$, $\mathcal{F}_E \cup \mathcal{G}_{\neg E} \in \mathcal{A}$

SAV0’ says that $\mathcal{A}$ contains not only all constant act-functions, but also all act-functions that can be constructed therefrom *via* a finite number of mixings. Note that, although $\mathcal{O}$ may contain an infinite number of outcomes, each act-function in $\mathcal{A}$ is only ever associated with a finite number of outcomes.

The first two real preference conditions are straightforward weak ordering and non-triviality requirements on $\succcurlyeq$:

SAV1 $\succcurlyeq$ on $\mathcal{A}$ is a weak ordering

SAV2 $\underline{o}_\mathrm{i} \succ \underline{o}_\mathrm{j}$ for some $o_\mathrm{i}, o_\mathrm{j} \in \mathcal{O}$

The transitivity of $\succcurlyeq$ is an obvious necessary condition for the kind of T-representation that Savage aims to achieve, whereas the completeness of $\succcurlyeq$ is required for Savage’s strong uniqueness result (amongst other things). **SAV2** is a simple non-triviality condition.

The remaining preference conditions require a bit of work to spell out. We first extend $\succcurlyeq$ to restricted act-functions:

Definition 5.9: $\succcurlyeq$ for restricted act-functions
$\mathcal{F}_E \succcurlyeq \mathcal{G}_E$ iff $\mathcal{F}^* \succcurlyeq \mathcal{G}^*$ whenever $\mathcal{F}_E = \mathcal{F}^*_E$, $\mathcal{G}_E = \mathcal{G}^*_E$, and $\mathcal{F}^*_{\neg E} = \mathcal{G}^*_{\neg E}$

Furthermore, define the set of *null events*, $\mathcal{N}$, as:

Definition 5.10: Null events
$\mathcal{N} = \{E \in \mathcal{E}\colon \mathcal{F} \sim \mathcal{G}$ whenever $\mathcal{F}_{\neg E} = \mathcal{G}_{\neg E}\}$

The members of $\mathcal{N}$ are the events which will receive a $\mathcal{B}el$ value of 0 in the final representation. Again, the idea behind this is highly intuitive: if any two act-functions are considered equivalent for the purposes of decision-making whenever they only differ in their outcomes with respect to states $s \in E$ for some event E, then what happens in those states must be considered utterly irrelevant from the point of view of the decision-maker. Assuming basic rationality, this would come to pass just in case the subject had zero confidence in one of those states obtaining. The background assumption, of course, is that agents have no interest the outcomes of their acts at states they consider utterly unlikely to be true.[54]

Savage's next two preference conditions express his so-called *sure-thing principle*. For all $\mathcal{F}$, $\mathcal{G}$, $\mathcal{F}^*$, $\mathcal{G}^* \in \boldsymbol{\mathcal{A}}$, $E \in \boldsymbol{\mathcal{E}}$, and $o_1, o_2 \in \boldsymbol{\mathcal{O}}$,

SAV3 If $\mathcal{F}_E = \mathcal{G}_E$, $\mathcal{F}^*_E = \mathcal{G}^*_E$, $\mathcal{F}_{\neg E} = \mathcal{F}^*_{\neg E}$, and $\mathcal{G}_{\neg E} = \mathcal{G}^*_{\neg E}$, then $\mathcal{F} \succ \mathcal{F}^*$ iff $\mathcal{G} \succ \mathcal{G}^*$

SAV4 If $E \in \boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{N}}$, then $\underline{o}_E \succ \underline{o}^*_E$ iff $\underline{o} \succ \underline{o}^*$

Savage's famous example of his principle goes as follows:

> A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he knew that the Democratic candidate were going to win, and decides that he would. Similarly, he considers whether he would buy if he knew that the Republican candidate were going to win, and again finds that he would. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event obtains, or will obtain, as we would ordinarily say. It is all too seldom that a decision can be arrived at on the basis of this principle, but except possibly for the assumption of simple ordering, I know of no other extralogical principle governing decisions that finds such ready acceptance. (1954, 21-2)

More specifically, **SAV3** says that whether $\succcurlyeq$ holds between two act-functions does not depend on those states which have identical consequences for the two acts. This seems plausible for any rational agent, given the assumption that the states are act-independent. **SAV4**, on the other hand, sets up a correspondence between outcome preferences (i.e., preferences over constant act-functions) and restricted act-function preferences for non-null events.

The next condition is especially important for the sensibility of Coherence. Say that $\mathcal{F}_E \equiv o$ iff $\mathcal{F}_E(s) = o$ for all s in E. Then, for all relevant acts and events,

SAV5 If $\underline{o}_1 \succ \underline{o}_2$, $\mathcal{F}_E \equiv o_1$, $\mathcal{F}_{\neg E} \equiv o_2$, $\mathcal{G}_{E^*} \equiv o_1$, $\mathcal{G}_{\neg E^*} \equiv o_2$, and similarly for $\underline{o}^+_1$, $\underline{o}^+_2$, $\mathcal{F}^+$, $\mathcal{G}^+$, then $\mathcal{F} \succ \mathcal{G}$ iff $\mathcal{F}^+ \succ \mathcal{G}^+$

This condition, in conjunction with Coherence, ensures that $\succcurlyeq^b$ is a weak ordering on $\boldsymbol{\mathcal{E}}$. To recall, Coherence tells us that a subject finds E_1 strictly more probable than E_2 just in case, for any pair of outcomes o_1 and o_2, whenever she prefers o_1 to o_2, she prefers the act-function $(E_1, o_1 \mid \neg E_1, o_2)$ over $(E_2, o_1 \mid \neg E_2, o_2)$. We interpret this as the one act having

[54] I will not delve into the plausibility of this assumption here, though I will note that it is not obviously true. See (Bradley and Stefansson forthcoming) for related discussion.

a higher subjective likelihood of resulting in the better outcome, and a lower likelihood of resulting in the worse outcome. **SAV5** says that *any* time a subject prefers (E_1, $o_1 \mid \neg E_1$, o_2) to (E_2, $o_1 \mid \neg E_2$, o_2) for *some* o_1, o_2 such that $o_1 \succ o_2$, then for *all* pairs of outcomes o_3, o_4 such that $o_3 \succ o_4$, the agent will prefer (E_1, $o_3 \mid \neg E_1$, o_4) to (E_2, $o_3 \mid \neg E_2$, o_4). In light of how we are interpreting the agent's behaviour, **SAV5** can be read as a basic condition of coherent decision-making upon an agent: if, in one instance, she is disposed to choose as if she considers E_1 more likely than E_2, then she ought to choose as such in all instances. Without this condition, a subject's preferences may fail to determine any well-defined qualitative probability relation at all, rendering Coherence effectively useless.

Savage's final two preference conditions are that, for all $\mathcal{F}$, $\mathcal{G} \in \boldsymbol{\mathcal{A}}$ and $E \in \boldsymbol{\mathcal{E}}$,

SAV6 If $\mathcal{F} \succ \mathcal{G}$ then there is a finite partition $\boldsymbol{T}$ of $\boldsymbol{\mathcal{S}}$ such that for all $E \in \boldsymbol{T}$, $\mathcal{F}^*_E \equiv o_1$ and $\mathcal{F}^*_{\neg E} = \mathcal{F}_{\neg E}$ only if $\mathcal{F}^* \succ \mathcal{G}$; and $\mathcal{G}^*_E \equiv o_1$ and $\mathcal{G}^*_{\neg E} = \mathcal{F}^*_{\neg E}$ only if $\mathcal{F} \succ \mathcal{G}^*$

SAV7 If $\mathcal{H}_E \equiv \mathcal{G}(s)$, then $\mathcal{F}_E \succ \mathcal{H}_E$ only if $\mathcal{F}_E \succcurlyeq \mathcal{G}_E$; and $\mathcal{H}_E \succ \mathcal{F}_E$ only if $\mathcal{G}_E \succcurlyeq \mathcal{F}_E$

SAV6 is a very strong structural condition which in effect requires that no outcome is either infinitely desirable or infinitely undesirable. In conjunction with the other preference conditions, it plays an important role in the derivation of a probability function $\mathcal{B}el$ that represents $\succcurlyeq^b$. **SAV7** is also very strong, but as noted above, it's not required if we limit our attention to finitely-valued act-functions.

With these conditions set out, Savage proves the following theorem:

Theorem 5.1: Savage's theorem

If **SAV0–SAV7** hold of <$\boldsymbol{\mathcal{S}}$, $\boldsymbol{\mathcal{E}}$, $\boldsymbol{\mathcal{N}}$, $\boldsymbol{\mathcal{O}}$, $\boldsymbol{\mathcal{A}}$, $\succcurlyeq$>, then there is a probability function $\mathcal{B}el$: $\boldsymbol{\mathcal{E}} \mapsto [0, 1]$, and a function $\mathcal{D}es$: $\boldsymbol{\mathcal{O}} \mapsto \mathbb{R}$, such that for all o_1, $o_2 \in \boldsymbol{\mathcal{O}}$, all E, E_1, $E_2 \in \boldsymbol{\mathcal{E}}$, and all ($E_\mathrm{i}$, $o_\mathrm{i} \mid \ldots \mid E_\mathrm{n}$, o_n), (E_j, $o_\mathrm{j} \mid \ldots \mid E_\mathrm{m}$, o_m) $\in \boldsymbol{\mathcal{A}}$,

(i) $\underline{o_1} \succcurlyeq \underline{o_2}$ iff $\mathcal{D}es(o_1) \geq \mathcal{D}es(o_2)$

(ii) $E_1 \succcurlyeq^b E_2$ iff $\mathcal{B}el(E_1) \geq \mathcal{B}el(E_2)$

(iii) If $0 < \lambda < 1$, then $\mathcal{B}el(E_1) = \lambda.\mathcal{B}el(E)$, for some $E_1 \subseteq E$

(iv) (E_i, $o_\mathrm{i} \mid \ldots \mid E_\mathrm{n}$, o_n) $\succcurlyeq$ (E_j, $o_\mathrm{j} \mid \ldots \mid E_\mathrm{m}$, o_m) iff $\sum_i^n \mathcal{B}el(E_\mathrm{i}).\mathcal{D}es(o_\mathrm{i}) \geq \sum_j^m \mathcal{B}el(E_\mathrm{j}).\mathcal{D}es(o_\mathrm{j})$

Furthermore, $\mathcal{B}el$ is unique and $\mathcal{D}es$ is bounded and unique up to positive linear transformation

A thorough statement of the proof of Theorem 5.1 can be found in (Fishburn 1970, Ch. 14).

The strong statement of Savage's uniqueness condition, while technically accurate, is somewhat misleading. Savage *does* prove that, *given a choice of* $\boldsymbol{\mathcal{O}}$ and $\boldsymbol{\mathcal{S}}$, if **SAV0–SAV7** are satisfied then $\succcurlyeq$ can be given an expected utility representation where $\mathcal{B}el$ on $\boldsymbol{\mathcal{E}}$ is unique and $\mathcal{D}es$ on $\mathcal{O}$ is unique up to positive linear transformation. The strength of this uniqueness condition is often considered a substantial point in favour of Savage's theorem. It's *prima facie* valuable to have a theorem which supplies us with a unique credence function. The problem here is that both $\mathcal{B}el$ and $\mathcal{D}es$ have their uniqueness conditions only relative to the choice of $\mathcal{S}$ and $\mathcal{O}$. This much is obvious for $\mathcal{D}es$, as it is a function defined on $\mathcal{O}$ and so necessarily changes its character whenever $\mathcal{O}$ is altered. But as Schervish, Seidenfeld *et al.* (1990) show, the character of $\mathcal{B}el$ is also strongly dependent not only on how $\mathcal{S}$ (and hence $\mathcal{E}$) is specified, but also on how $\mathcal{O}$ is specified: if it turns out that there are multiple, equally viable ways of characterising the space of states and outcomes, then Savage's strong uniqueness results are to some extent illusory (see also Levi 2000, 399).

A huge number of decision-theoretic representation theorems are formulated within a framework very similar to Savage's own. As Krantz *et al.* put it in their monumental *Foundations of Measurement*,

> In general, a rough sort of consensus exists about the primitive terms to be employed in the formulation of the problem of decision making under risk or uncertainty. Nearly everyone seems to agree that there are chance events to which probabilities adhere, consequences which exhibit utilities, and decisions that are more or less arbitrary associations of consequences to events. (1971, 411)

That is to say, a great many representation theorems (then and today) begin with $\mathcal{S}$ and $\mathcal{O}$, and define $\succcurlyeq$ on a collection $\mathcal{A}$ of act-functions. Most theorists working within the paradigm Savage created define $\mathcal{A}$ as the set of *all* total functions from $\mathcal{S}$ to $\mathcal{O}$. Others have taken $\succcurlyeq$ to be defined on only a proper subset of $\mathcal{O}^{\mathcal{S}}$ (e.g., Richter 1975, Wakker and Zank 1999, Casadesus-Masanell, Klibanoff *et al.* 2000), or on partial functions from $\mathcal{S}$ to $\mathcal{O}$ (e.g., Luce and Krantz 1971, Luce 1972, Roberts 1974, Narens 1976).[55]

Importantly, these theorems include not only those for classical expected utility theory, but a very wide range of non-expected utility theories as well. Indeed, the vast majority of NCU theorems belong to the Savage paradigm. In Appendix B, I outline four distinct NCU theorems formulated using Savage's formal framework, though these four only scratch the surface. Savage's own theorem, as a CEU theorem, is limited to probabilistic $\mathcal{B}el$ functions. On the other hand, the huge variety of representation styles that can and have been arrived at through the use of Savage's framework—many of which allow for

[55] Suppes (1969) and Fishburn (1967) diverge from the general trend by characterising their basic objects of preference as ordered pairs of Savage-style act-functions (i.e., the option space is a subset of $\mathcal{O}^{\mathcal{S}} \times \mathcal{O}^{\mathcal{S}}$), which are supposed to represent even-chance bets with the performances of different acts as prizes. The theorem of (Kochov 2015) has a rather unique formal structure, but its basic relata for $\succcurlyeq$ can be accurately described as "multiperiod counterparts of Savage act[-function]s" (240).

non-probabilistic $\mathcal{B}el$ functions—should be encouraging to proponents of characterisational representationism. Unfortunately, though, there are a number of issues that arise from the use of the framework itself, to which we now turn.

5.2 Constant act-functions and imaginary acts

I will begin my critical discussion with what is easily the most frequently cited objection to Savage's system, which Fishburn (1981) calls the *constant acts problem*: it's implausible that constant act-functions can serve to represent anything that an ordinary agent could choose to *do*. If $\mathcal{A}$ is supposed to represent the space of acts available to the agent in her current situation, then constant act-functions are an anomaly—functions which represent nothing in the real world that the agent could have preferences between.

Fishburn (1981) gives the following illustration of the problem. Let the outcome o be *Carrying an umbrella on a bright and sunny day*, and the event E be *It rains*. Then, every s in E is a state in which it rains, and any act-function which maps an s in E to o is pairing an outcome with a state that is inconsistent with it. "In fact, the natural set of [outcomes] that could occur under one state may be disjoint from the set that could occur under another state" (1981, 162). Note that, on this way of describing the issue, the problem appears to be that constant act-functions may pair outcomes with incompatible states, thus apparently representing acts which are literally impossible to perform.[56] If s and o are logically inconsistent, then not even an omnipotent god could make it the case that s and o. Suppes and Luce (1965, 299), Karni (1993), and Maher (1993, 182-5) give a similar account of the constant acts problem as involving inconsistent state and outcome pairings.[57]

However, the issues here are somewhat more subtle than they are often made out to be. Constant act-functions do give rise to difficulties for characterisational representationism, but exactly what these difficulties may be depends on how we interpret the relevant formalisms. Let us therefore look again in depth at the origins of the constant acts problem, before we turn to how the problem might be dealt with.

[56] Indeed, if we make our outcomes so fine-grained that each outcome entails a conjunction of the form (*s obtains and* α *was performed*), as some are wont to do, then *every* finite-valued act-function in Savage's system will pair at least one outcome with an incompatible state.

[57] Joyce (1999, 107-8) also supposes that **SAV0** implies the existence of act-functions which pair together incompatible states and outcomes, but interprets the constant acts problem as arising primarily from the conjunction of the completeness requirement (entailed by **SAV1**) and **SAV0**. This is because he drops the behavioural interpretation of $\succcurlyeq$ for another interpretation compatible with preferences over non-existent acts. See §5.2.3.

5.2.1 The basis of the problem

The complaint about constant act-functions is usually levelled at **SAV0**, or its weaker counterpart **SAV0'**, wherein the character of $\mathcal{A}$ is formally specified. However, we must be careful not to lay all the blame on Savage's act-richness assumption—it is part of the problem, of course, but it's not the whole story. In fact, there are three independent factors which together lead to the constant acts problem, as I will now argue.

If taken purely on their own, **SAV0** and **SAV0'** are hardly problematic—each merely characterises $\mathcal{A}$ as a subset of $\mathcal{O}^{\mathcal{S}}$. What **SAV0**/**SAV0'** can be taken to require *in context* therefore depends on how the states in $\mathcal{S}$ are characterised, how the outcomes in $\mathcal{O}$ are characterised, and what the act-functions in $\mathcal{A}$ are intended to represent. Let us begin the interpretation of $\mathcal{A}$. As in §5.1.1, we will assume that every act-function is assumed to correspond to something an agent might *do*. Let us call this the *Act–Function Correspondence* assumption, which can be stated as follows:

> **Act–Function Correspondence**
> There exists a natural, one-one correspondence between the set of act-functions $\mathcal{A} \subseteq \mathcal{O}^{\mathcal{S}}$ and the space of available acts $\mathcal{A}'$ such that every $(E_1, o_1 | \ldots | E_n, o_n) \in \mathcal{A}$ represents a unique act (or set of acts with the same pattern of consequences) in A' which, if performed, would result in o_1, if any $s \in E_1$ were the case, …, and o_n if any $s \in E_n$ were the case

As we've seen, Act–Function Correspondence requires that states are at least *logically* act-independent, and outcome-functional; if states did not have these properties, the representation of acts using act-functions would make little sense.

SAV0/**SAV0'** and Act–Function Correspondence are not yet enough to get us a problem—we still need to specify the nature of the outcomes. To see this, note that it's consistent with Savage's formalism that the outcomes in $\mathcal{O}$ are very coarse-grained. Suppose, then, that $\mathcal{O}$ contains only two extremely non-specific outcomes, o_1 and o_2. For instance, let o_1 and o_2 be very long, mutually exclusive disjunctions of the more specific states of affairs that we would ordinarily consider the outcomes of a decision to be. In this case, there does not appear to be anything unusual about constant act-functions: $\underline{o}_1$ and $\underline{o}_2$ could be construed simply as acts (or a collection of acts) which result in one or another disjunct becoming true—and such 'acts' are ubiquitous. The problem with this, of course, is that characterising $\mathcal{O}$ this way conflicts with the informal requirement of context neutrality—without which Savage's preference conditions and his principle of Coherence become highly implausible. For similar reasons, we can assume that any *useful* representation of acts as functions from $\mathcal{S}$ to $\mathcal{O}$ should make use of rather fine-grained outcomes.

We now have enough for the constant acts problem to arise. Generally speaking, there is a deep tension within Savage's system between the following triad:

(1) **SAV0**/**SAV0’**

(2) Act–Function Correspondence

(3) Fine-grained outcomes

A theorist could reasonably pick any two of these to adopt, but trying to justify all three at once is difficult. Let us assume (3) in all that follows. In this case, the constant acts problem becomes clear: **SAV0**/**SAV0’** implies that $\mathcal{A}$ has a particular kind of formal structure; Act–Function Correspondence in turn requires that $\mathcal{A}$’ must have the same structure. The existence of constant act-functions in $\mathcal{A}$, however, seems incompatible with Act–Function Correspondence. One of these needs to go.

There are two lessons that I wish to draw here. The first is that it is *slightly* misleading to express the problem as being about the *compatibility* of some states and outcomes. There would still be cause to worry about Act–Function Correspondence even if there were no mutually incompatible pairs of states and outcomes, and the problematic act-functions are by no means limited only to those which pair together incompatible states and outcomes. On *any* natural conception of acts and outcomes, immensely implausible that there is an act we can perform such that, *regardless of how the world turns out to be independently of our decision*, one and only one fine-grained outcome will obtain. Now, this *may* be because the set of potential outcomes $\mathcal{O}_1$, $\mathcal{O}_2 \subseteq \mathcal{O}$ that may result from any available act at two distinct states s_1 and s_2 respectively only partially overlap, if they overlap at all—indeed, this would seem to be the so in any ordinary case: some states just don’t play nicely with some outcomes. However, even supposing that *every* state is consistent with the same range of outcomes, there would still be no good reason to think that $\mathcal{A}$’ has the kind of structure imposed upon it by the conjunction of **SAV0**/**SAV0’** and Act–Function Correspondence. Which outcomes can arise in which states depends on the range of acts available to the agent at the time of the decision, and **SAV0**/**SAV0’** places rather implausible constraints on what that range of acts must always look like. *The* problem, therefore, is not simply that:

> In virtually any realistic problem that is formulated in the Savage mode, some consequences will be incompatible with some states or events, as is “carry an umbrella on a bright, sunny day” with “rain”. (Fishburn 1981, 162)

Rather, the problem is the unjustified and implausible imposition of a particular structure upon $\mathcal{A}$’.[58]

[58] In Fishburn’s example, *It rains* is an event—but given an outcome set $\mathcal{O}$ that includes *Carrying an umbrella on a bright and sunny day*, there cannot be any such event *in* $\mathcal{E}$. As noted in §5.1.1, states must be act-independent, outcome-functional, and thus, in light of **SAV0**/**SAV0’** and Act–Function Correspondence, events must be outcome-independent. Of course, rain could still *occur*—the point is that there can be

The second thing to note is that constant act-functions are only a very small part of a broader problem. For example, essentially the same worries that arise for constant act-functions can be raised for what we might call bifurcating act-functions, or act-functions of the form $(E, o_1 | \neg E, o_2)$; and likewise for trifurcating act-functions $(E_1, o_1 | E_2, o_2 | E_3, o_3)$, and so on. Most (if not all) act-functions which range over only a small number of distinct finely-individuated outcomes will be just as problematic as constant act-functions, and for essentially the same reasons. I will refer to any act-function which lacks a corresponding act in $\mathcal{A}'$ as an *imaginary act-function*.[59] Any imaginary act-function causes as much trouble for Savage as a constant act-function does—the constant functions are simply the most salient example of the underlying issue.

If one wants to avoid the bigger issues at the heart of the constant acts problem, it is clear that one must do much more than just remove constant act-functions from $\mathcal{A}$. The presence of *imaginary* act-functions in $\mathcal{A}$ is problematic *inasmuch* as $\mathcal{A}$ is supposed to represent $\mathcal{A}'$. This leaves us with two options. On the one hand, one might retain Act–Function Correspondence and try to develop a theorem around a more realistic representation of $\mathcal{A}'$. On the other hand, one could drop Act–Function Correspondence, offering instead an alternative interpretation of the system which somehow makes sense of imaginary act-functions. In the remainder of this section, I will consider the viability of each of these options in turn.

5.2.2 Doing without imaginary act-functions

Given Act–Function Correspondence, **SAV0** and even the weaker **SAV0'** are clearly too strong. For characterisational representationism, this will not do. The problem here is not just that *ordinary* agents could not have preferences satisfying the conditions, but rather that it would not even make sense to assert of *anyone* that their preferences satisfy the conditions. To say that these act-richness assumptions are *false* is to say that $\succcurlyeq$ is formally required to have a domain which it *does not*, and in fact *cannot*, have (and thus brings the theorem into conflict with desideratum **(1a)**).

Some have thought to respond to the problem of constant acts by weakening those act-richness assumptions. As noted earlier, Luce and Krantz (1971) were able to obtain a

no event *in* $\mathcal{E}$ which corresponds to that proposition *if Carrying an umbrella on a bright and sunny day* already exists in $\mathcal{O}$. To apply Savage's system, we are not free to pick and choose as we like our states, outcomes, and events, but must do so within tightly constrained limits. As I will argue below, this fact itself leads to further problems with Savage's framework.

[59] Maher (1993, 183) refers to these as *uninterpretable acts*.

representation result without requiring the use of constant act-functions, which they consider an important benefit of their approach.[60] However, we have seen that simply removing constant act-functions from $\mathcal{A}$ is inadequate as a response to the broader problem with imaginary act-functions. Luce and Krantz retain still very strong assumptions about the structure of their set of act-functions (see Appendix B), which by their own admission seem to imply the presence of imaginary act-functions. This is the basis of Joyce's (1999, 108-10) critique of Luce and Krantz's theorem, and I will not add anything further to it here.[61]

There is a general reason for this failure: like Savage, Luce and Krantz attempt to formally construct their set of act-functions $\mathcal{A}$ using just $\mathcal{S}$ and $\mathcal{O}$ but independently of any knowledge or specifications regarding the space of available acts $\mathcal{A}'$. It is unreasonable to begin with an arbitrary partition $\mathcal{S}$ and an equally arbitrary set of outcomes $\mathcal{O}$, and expect to work backwards from there to arrive at a plausible reconstruction of the space of available acts. $\mathcal{A}'$ may correspond to a proper subset of some collection of act-functions (defined for *some* ways of construing $\mathcal{S}$ and $\mathcal{O}$), but the *formal* character of this subset will depend heavily on the nature of $\mathcal{A}'$ itself. There may, for instance, be one available act $(E, o_1 | \neg E, o_2)$ but no $(E, o_2 | \neg E, o_1)$, or vice versa—but there is no way to know this, if all that is given is $\mathcal{S}$ and $\mathcal{O}$. If Act–Function Correspondence is ever to be justified, the *formal* construction of the space of act-functions needs to begin with $\mathcal{A}'$.[62]

On the flip side, however, as I will now argue, it is very difficult (if not impossible) to develop a Savage-like representation theorem *without* making some rather strong, and ultimately implausible, assumptions about $\mathcal{A}$. There are multiple reasons for this, though in what follows I will focus upon what appears to me the most troubling: the difficulty in developing well-defined orderings on $\mathcal{E}$ and $\mathcal{O}$, needed to construct $\mathcal{B}el$ and $\mathcal{D}es$ respectively, without an appeal to imaginary act-functions.

Fishburn has argued that, without appealing to constant act-functions, "there is no natural way of defining preferences on [outcomes] in terms of preferences on acts" (1970,

[60] See also (Gaifman and Liu 2015) for a recent attempt at *minimising*—but not altogether removing—the use of constant act-functions within a Savagean framework. Gaifman and Liu's theorem requires that there are at least two constant act-functions. Although much weaker than **SAV0**, it's not at all clear that their replacement condition (or the more general assumptions they need to make about the structure of their set of act-functions) is consistent with Act–Function Correspondence.

[61] A further problem with Luce and Krantz's formalisation is that many of their act-functions are very difficult to interpret as acts. See (Krantz and Luce 1974), (Spohn 1977), and (Fishburn 1981) for discussion.

[62] I am unaware of any Savage-like theorems which take the path I am suggesting, though it is briefly discussed by Fishburn (1970, 164-7). Balch and Fishburn (1974, see also Balch 1974, Fishburn 1974) develop a theorem which begins with a primitive set of acts $\mathcal{A}'$ and a set of act-independent states $\mathcal{S}$, with outcomes defined as act-event pairs. Their theorem belongs to the class of lottery-based theorems, which are discussed below.

166). In Savage's system, however, preferences over constant act-functions form a crucial part of constructing the $\mathcal{D}es$ function—recall that, in his representation,

$\underline{o}_1 \succcurlyeq \underline{o}_2$ iff $\mathcal{D}es(o_1) \geq \mathcal{D}es(o_2)$

Thus, Fishburn suggests that to do without constant act-functions, a theorist would need to develop a dual-primitive theorem, with $\succcurlyeq$ defined on $\boldsymbol{\mathcal{A}}$ and a separate preference relation $\succcurlyeq^{u}$ defined on $\boldsymbol{\mathcal{O}}$. As it turns out, though, with some imagination it is possible to characterise relative utilities between outcomes in terms of preferences between act-functions without appealing to constant act-functions at all. It will be instructive to see why this alternative characterisation still seems to end up requiring an appeal to imaginary act-functions.

The basic idea here is dominance reasoning: an outcome o_1 is more desirable than another outcome o_2 for an agent S iff $\mathcal{F} \succ \mathcal{G}$, when $\mathcal{F}$ and $\mathcal{G}$ only differ, *with respect to the states that S gives some credence to*, in that $\mathcal{F}$ is sometimes paired with o_1 at some states while $\mathcal{G}$ is paired with o_2 at those same states. In this case, with respect to what the agent considers possible, $\mathcal{F}$ represents an act which is identical to the act represented by $\mathcal{G}$ but for the possibility of resulting in o_1 instead of o_2 at some states—and if $\mathcal{F} \succ \mathcal{G}$, this is presumably then because o_1 is preferred to o_2.

In order to spell this idea out formally, we will first need a notion of *nullity* for states. As a consequence of Definition 5.10, any subset of a null event is also null, including any singleton events $\{s\}$, for $s \in E \in \boldsymbol{\mathcal{N}}$. Given this, say that a state is *null* iff it belongs to an event E and E is null in the sense of Definition 5.10; the state is *non-null* otherwise. Now let $\boldsymbol{\mathcal{S}}' \subset \boldsymbol{\mathcal{S}}$ be a set of non-null states. We can now define a relative utility ranking $\succcurlyeq^{u}$ as follows:

Definition 5.11: $\succcurlyeq^{u}$ without constant acts

$o_1 \succcurlyeq^{u} o_2$ iff $\mathcal{F} \succcurlyeq \mathcal{G}$ whenever, for some set of non-null states $\boldsymbol{\mathcal{S}}'$,

(i) If $s \in \boldsymbol{\mathcal{S}}'$, then $\mathcal{F}(s) = o_1$ and $\mathcal{G}(s) = o_2$

(ii) For all non-null $s \notin \boldsymbol{\mathcal{S}}'$, $\mathcal{F}(s) = \mathcal{G}(s)$

Assuming that outcomes are context neutral, the right-to-left direction of Definition 5.11 seems plausible for any rational agent—the dominance principle it embodies is one of the most intuitive precepts of folk decision theory. Furthermore, this definition does away with any need for constant act-functions.

However, there seems to be no good reason to think that the space of available acts will have the structure required for the general applicability of Definition 5.11. There are two distinct issues here.[63] The first arises as a result of the appeal to Definition 5.10 in the definition of null states. As almost any event in $\mathcal{E}$ can be null, and because we cannot presume to know *a priori* what events the agent considers null or non-null, the general application of Definition 5.10 already imposes quite strong restrictions upon the character of $\mathcal{A}$. That is, for any *potentially* null event E, Definition 5.10 requires that we will be able to find at least two act-functions which differ for some state(s) in E but which are identical with respect to all states in $\neg E$. There is no good reason to suppose that such acts will always be available.

Now, perhaps this first issue could be solved using another definition of nullity; or, alternatively, we might even assume that $\mathcal{N}$ is given to us for free as a primitive. This will not be enough, because a closely related issue arises for Definition 5.11 itself. In particular, in order to ensure that the left-to-right direction always holds for any potential subject S, it will need to be the case that for *every* way of dividing the null states from the non-null there must be act-functions $\mathcal{F}$ and $\mathcal{G}$ which satisfy the stated conditions (i) and (ii) with respect to the relevant outcomes. This is still too strong an assumption, and there is no guarantee that the space of available acts will play along. An obvious example for when Definition 5.11 cannot be applied (but certainly not the only one) is the case of a fatalist who is certain that whatever outcome may eventually obtain, it will obtain regardless of her choices. At every state, she believes, any of her acts will result in the same outcome, whatever that outcome may be. The fatalist prefers some outcomes over others, and is uncertain about which outcome will obtain, but there will be *no* acts available to her which have different outcomes at any states *she gives credence to*; hence, any act-function which satisfies (i) is imaginary.

Suppose, then, that both $\mathcal{N}$ and $\succcurlyeq^{\mathrm{u}}$ are given as primitives, *not* defined in terms of preferences on act-functions. There is now the problem of defining $\succcurlyeq^{\mathrm{b}}$, needed to construct the $\mathcal{B}el$ function, without making undue assumptions about the character of $\mathcal{A}$'. Savage's principle of Coherence appeals to bifurcate act-functions, which are usually no more plausible *qua* representations of available acts than constant act-functions. So, an alternative definition for $\succcurlyeq^{\mathrm{b}}$ will need to be found as well.

Machina and Schmeidler (1992) present a somewhat more plausible definition of $\succcurlyeq^{\mathrm{b}}$ within an essentially Savagean framework, as follows:

Definition 5.12: $\succcurlyeq^{\mathrm{b}}$ (Machina and Schmeidler)

$E_1 \succcurlyeq^{\mathrm{b}} E_2$ iff, if $o_1 \succ^{\mathrm{u}} o_2$, then $\mathcal{F} \succcurlyeq \mathcal{G}$ whenever:

(i) If $s \in E_1 - E_2$, then $\mathcal{F}(s) = o_1$ and $\mathcal{G}(s) = o_2$

(ii) If $s \in E_2 - E_1$, then $\mathcal{F}(s) = o_2$ and $\mathcal{G}(s) = o_1$

(iii) If $s \notin (E_1 - E_2) \cup (E_2 - E_1)$, then $\mathcal{F}(s) = \mathcal{G}(s)$

[63] To focus in on the main problem, I will assume for now that $\succcurlyeq$ is complete on $\mathcal{A}$; in §5.2.4 I will discuss what can be said when that assumption is false.

The reasoning behind Definition 5.12 is very similar to the reasoning behind Coherence. Indeed, the two definitions amount to the same thing in the special case where $E_2 = \neg E_1$. If $\mathcal{F}$ and $\mathcal{G}$ satisfy the stated conditions, then the agent would prefer $\mathcal{F}$ to $\mathcal{G}$ iff she found E_1 more likely than E_2, as $\mathcal{F}$ has the greater subjective likelihood of resulting in the better outcome. The major benefit of Machina and Schmeidler's definition is that it does not make use of bifurcate act-functions—in fact, $\mathcal{F}$ and $\mathcal{G}$ may have any number of outcomes. Unfortunately, Machina and Schmeidler's alternative still imposes strong constraints on the space of available acts. Before I argue this, however, I will note that it's possible to improve upon their definition in at least three ways.

To begin with, the reasoning which underlies the definition does not require something as strong as condition (ii), which makes mention of the same *outcomes* as appeared in condition (i). It would be enough that the second condition appeals to outcomes with the same *utilities* as those mentioned in (i); and since we have taken $\succcurlyeq^{\mathrm{u}}$ as a primitive we can replace (ii) with:

(ii') If $s \in E_2 - E_1$, then $\mathcal{F}(s) = o_4$ and $\mathcal{G}(s) = o_3$, where $o_3 \sim^{\mathrm{u}} o_1$, and $o_4 \sim^{\mathrm{u}} o_2$

The outcome o_3 may or may not be identical to o_1, and similarly for o_2 and o_4, so (ii') is a strictly weaker condition than (ii). The second improvement is similar: with respect to condition (iii), sameness of outcomes is unnecessary—sameness of utility would be enough. (Strictly, it would be enough that the credence-weighted average of the outcomes under the states $s \notin (E_1 \cup E_2)$ is equal for $\mathcal{F}$ and $\mathcal{G}$, but there is no obvious way to specify such a condition prior to deriving the credence function.) Thus we can replace (iii) with:

(iii') If $s \notin (E_1 - E_2) \cup (E_2 - E_1)$, then $\mathcal{F}(s) \sim^{\mathrm{u}} \mathcal{G}(s)$

Finally, it's possible to weaken the definition's requirements on $\boldsymbol{\mathcal{A}}$ if all null events are discounted from consideration. Definition 5.12 applies to *all* pairs of events E_1 and E_2, and so act-functions must be found which satisfy the definitions three conditions with respect to any pair E_1 and E_2. However, null events can be presumed to sit at the bottom of the $\succcurlyeq^{\mathrm{b}}$ ranking (to be assigned a credence of 0), so we don't need to consider preferences over act-functions to decide where they sit with respect to $\succcurlyeq^{\mathrm{b}}$.

The foregoing then leads to the following, improved definition of $\succcurlyeq^{\mathrm{b}}$:

Definition 5.13: $\succcurlyeq^{\mathrm{b}}$ (Machina and Schmeidler improved)
If $E \in \boldsymbol{\mathcal{N}}$, then for all $E' \in \boldsymbol{\mathcal{E}}$, $E' \succcurlyeq^{\mathrm{b}} E$; and for all $E_1, E_2 \in \boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{N}}$, $E_1 \succcurlyeq^{\mathrm{b}} E_2$ iff, if $o_1 \succ^{\mathrm{u}} o_2$, then $\mathcal{F} \succcurlyeq \mathcal{G}$ whenever

(i) If $s \in E_1 - E_2$, then $\mathcal{F}(s) = o_1$ and $\mathcal{G}(s) = o_2$

(ii') If $s \in E_2 - E_1$, then $\mathcal{F}(s) = o_4$ and $\mathcal{G}(s) = o_3$, where $o_3 \sim^{\mathrm{u}} o_1$, and $o_4 \sim^{\mathrm{u}} o_2$

(iii’) If $s \notin (E_1 - E_2) \cup (E_2 - E_1)$, then $\mathcal{F}(s) \sim^{u} \mathcal{G}(s)$

The justification for Definition 5.13 is essentially identical to the justifications for Definition 5.12 and Coherence, but it places strictly weaker requirements on the structure of $\mathcal{A}$ than either of the latter two definitions.

It will come as no surprise that Definition 5.13 is still too strong. To *ensure* that $\succcurlyeq^{b}$ is always well-defined, it must be assumed that there will always be some $\mathcal{F}$ and $\mathcal{G}$ satisfying the conditions (i), (ii’), and (iii’), for any pair of non-null events E_1 and E_2 that we care to choose. And there are good reasons to think that this will not always be the case. Here is a schematic example.[64] Let E_1 be an event where, independently of any acts I might perform, many very good things occur, and let E_2 be an event where a great deal of very horrible things occur independently of any act I might perform. For simplicity, suppose that E_1 and E_2 are disjoint events. In fact, suppose that E_1 is so much better than E_2 that the very *best* possible outcome that might obtain if E_2 were true would still be worse than the very *worst* outcome that might obtain given E_1. If this is the case, however, then any act-function which satisfies (i) and (ii’) *cannot* represent an available act: there *are* no acts α and β, for instance, such that α leads to o_1 at E_1, and β leads to $o_3 \sim o_1$ at E_2. According to Definition 5.13 then, E_1 and E_2 are *incomparable* with respect to $\succcurlyeq^{b}$.

A final illustration of the difficulties that come with trying to remove imaginary act-functions should suffice. As it turns out, there does appear to be a way to *systematically* construct a set of act-functions from a set of states and outcomes so as to *guarantee* act-independence, outcome-functionality, and Act–Function Correspondence. The strategy is based on a discussion of Lewis’ (1981); Gibbard and Harper (1978) and Stalnaker (1972) also refer to a closely related idea, and it’s critically discussed by Joyce (1999, 115-19). First of all, take $\mathcal{A}$’—that is, a set of *acts* rather than act-functions—and $\mathcal{O}$ as primitive. It is assumed that the outcomes in $\mathcal{O}$ are mutually exclusive and jointly exhaustive, consistent with the performance of any act in $\mathcal{A}$’, and context neutral. $\mathcal{S}$ can now be defined as the set of all functions from $\mathcal{A}$’ to $\mathcal{O}$.

For instance, suppose there are only two available acts, α and β, and only two possible outcomes, o_1 and o_2. Then $\mathcal{S}$ contains four distinct functions:

$s_1 = \{(\alpha, o_1), (\beta, o_1)\}$
$s_2 = \{(\alpha, o_1), (\beta, o_2)\}$
$s_3 = \{(\alpha, o_2), (\beta, o_1)\}$
$s_4 = \{(\alpha, o_2), (\beta, o_2)\}$

[64] Thanks to Rachael Briggs for discussion here, and for help with this example. Exactly the same example also shows that Definition 5.12 and Coherence cannot always be applied.

In Lewis' terminology (1981, 11), each $s \in \mathcal{S}$ can be taken to represent a *dependency hypothesis*; i.e., a conjunction of counterfactuals which describes one of the different possible ways that the outcomes in $\mathcal{O}$ could causally depend upon the acts the agent might perform. For instance, s_1 can be read as *Regardless of what I do, o_1 obtains*, while s_2 is *If I do* α*, then* o_1 *will result, but if I do* β*, then* o_2 *will result*. Every dependency hypothesis is then (causally and hence logically) act-independent and outcome-functional (but *not* outcome-independent). Furthermore, given our assumptions, the set of dependency hypotheses is a partition of the relevant logical space.

With this in hand, each act in $\mathcal{A}'$ can be paired directly with an act-function in $\mathcal{A} \subset \mathcal{O}^{\mathcal{S}}$:

$$\alpha \triangleq \mathcal{F} = \{(s_1, o_1), (s_2, o_1), (s_3, o_2), (s_4, o_2)\}$$
$$\beta \triangleq \mathcal{G} = \{(s_1, o_1), (s_2, o_2), (s_3, o_1), (s_4, o_2)\}$$

The construction is such that there are *never* any constant act-functions. On the other hand, there will be *constant states*, or dependency hypotheses which imply that every act results in the same outcome. A consequence of these constant states is that the range of *every* act-function includes the entirety of $\mathcal{O}$. Moreover (as evidenced in the given example), act-functions will always evenly distribute the outcomes in $\mathcal{O}$ amongst the states in $\mathcal{S}$. For example, if there are 3 outcomes and 4 available acts, and thus $3^4 = 81$ states, each act-function will distribute each of the three outcomes to exactly 27 of those states. Thus, if there are more than 2 outcomes, we will never find *bifurcating* acts in $\mathcal{A}$ either (which figure centrally in Coherence).

Because $\mathcal{A}'$ is taken as primitive, and $\mathcal{A}$ is ultimately defined in terms of it, Act–Function Correspondence can hardly be doubted on this picture—indeed it seems about as plausible as it possibly can be. However, it also evident that none of the suggested definitions of $\succcurlyeq^{u}$ and $\succcurlyeq^{b}$ discussed above will be adequate if we adopt this framework. The Lewisian set of act-functions $\mathcal{A}$ has an interesting, and mathematically very elegant, structure to it—but it's the wrong kind of structure to guarantee that $\mathcal{N}$, $\succcurlyeq^{u}$, and $\succcurlyeq^{b}$ will always, or even *often*, be defined if Coherence, Definition 5.10, Definition 5.11, and/or Definition 5.13 are adopted. For example, the existence of constant states is enough to ensure that the earlier example given against Definition 5.13 applies; and Definition 5.11 cannot usefully be applied to any fatalist whose credence is distributed only over constant states. It may, of course, be possible to develop an interesting representation theorem based on this kind of construction—though I don't see how—but whatever it may turn out to be like, it will be quite different in its construction of $\mathcal{B}el$ and $\mathcal{D}es$ than anything Savage or his followers have put forward.

All of this suggests that it's very difficult—at best—to construct a Savage-like representation theorem without making some very strong assumptions about the set of act-functions, which seem implausible if Act–Function Correspondence is assumed. Savage's definitions of $\succcurlyeq^{u}$ and $\succcurlyeq^{b}$ are obviously off the table, but so are nearby suggestions. This point is borne out by other representation theorems developed within the Savage paradigm. These theorems typically require, if not constant acts, then at least a very richly structured $\mathcal{A}$ involving some imaginary act-functions. It would be an interesting project to see whether any interesting result can be achieved using the dependency hypothesis framework, but for the purposes of this discussion the key point is that no such results have been discovered—nor is it obvious than any will be found.

The presence of imaginary act-functions in $\mathcal{A}$ and Act–Function Correspondence are jointly inconsistent. So far, I have considered removing imaginary act-functions from the picture. I have argued that it seems highly unlikely that a Savage-like representation theorem will be developed under which Act–Function Correspondence is plausible. Nevertheless, removing imaginary act-functions from $\mathcal{A}$ is not the only possible response to the constant acts problem. Many authors working within the Savage paradigm are content to define $\succcurlyeq$ over imaginary act-functions, and *ipso facto* reject Act–Function Correspondence. It is to that response that I now turn.

5.2.3 Imaginary acts and (im)possible patterns of outcomes

Savage did not publish a response to the constant acts problem, though Fishburn (1981, 162-3) reports that it "did not greatly bother Savage since he felt that the preference comparisons required by his axioms were conceptually reasonable". Exactly what Fishburn meant by this is unclear, but many have taken it to mean that Savage was content to deal with preferences over *imaginary acts*—acts which, while not actually available for the agent to perform, could still in some sense or other be imagined.[65] Others—perhaps even most who have applied the Savage framework—have expressed similar sentiments.[66] That is, the most common response to the constant acts problem is that it seems conceptually possible to *imagine* some act which gives rise to such-and-such outcomes dependent on such-and-such states of the world obtaining, even if it's granted that the outcomes might be inconsistent with the states.

[65] See, e.g., (Levi 2000, 398): "Savage's approach does not require that the preference ranking over potential options be a preference ranking over actual options … There is textual evidence that Savage clearly understood this." I think Levi is entirely right about this—in particular, if constant act-functions are understood as representing genuinely available acts, then decision theory becomes trivial: every agent ought to perform the constant act which results in the best possible outcome at any state (Joyce 1999). Since he obviously did not intend for his theory to be trivial, it's plausible that Savage took some of his act-functions to represent imaginary acts. However, there is also textual evidence that Savage did not fully appreciate what this meant for his supposedly 'behaviouristic' definition of credences and utilities, and it conflicts sharply with how he introduces his decision theory in the early pages of his (1954).

[66] See (Buchak 2013, 91-2) for a recent example.

Unfortunately, it is very rare that much more is said on the issue beyond the bare assertion that imaginary acts make sense and that we can have preferences over such things. This situation is unsatisfactory; as I have been stressing, the interpretation of any one element of Savage's formalism is intimately tied up with the interpretation of every other element, and the introduction of imaginary acts into the intended interpretation of $\mathcal{A}$ has important consequences elsewhere. Most importantly, the inclusion of imaginary acts is incompatible with Savage's proposed interpretation of $\succcurlyeq$: "Loosely speaking, [α] $\succcurlyeq$ [β] means that, if [the subject] were required to decide between α and β, no other acts being available, he would decide on α". It is hard to make sense of this behavioural interpretation as being even "loosely" adequate if α and/or β are imaginary acts, especially if they are acts which result in inconsistent state-outcome pairs.

Preferences between imaginary acts call for a non-behavioural construal of $\succcurlyeq$, and it's evident in the literature that those who adopt imaginary acts as part of their interpretation of Savage's act-functions forego the behavioural reading of $\succcurlyeq$ in favour of a somewhat more mentalistic construal. Indeed, Broome (1991, 1993) refers to preferences over imaginary acts as *non-practical preferences*, as whatever preferences they represent cannot be manifest in agents' dispositions to choose between available acts. And James Dreier describes the self-elicitation of non-practical preferences as follows:

> Asked whether I prefer [α] or [β], I imagine myself in a situation in which I have to choose between them. I find myself inclined to choose [α]. I report, on that basis, that I prefer [α] to [β]. (1996, 268)

Supposing that every act-function corresponds to some imaginable act, one could interpret $\succcurlyeq$ as encoding an agent's dispositions to *judge* that she would choose one imagined act over another. Sobel's (1997) notion of a 'pairwise preference' is described in a similar vein.

It is somewhat doubtful that we can always conceive of an act which corresponds to an arbitrarily chosen pattern of outcomes—I at least struggle to picture an act which always brings it about that, say, *I have a glass of iced tea*, even at worlds where tea does not exist. There is, however, perhaps a more reasonable way to understand the situation, suggested by the following passage by Glen Shafer:

> [Savage] saw no reason why a person could not think about patterns of consequences corresponding to imaginary acts and formulate preferences between such patterns. In order to construct a preference between one pattern of consequences and another, it is not necessary that a person should have available a concrete act that produces this pattern, *or even that the person should be able to imagine such an act*. (1986, 470, emphasis added)

Instead of representing *acts*—whether real or imagined—by virtue of describing their patterns of outcomes, we might instead suppose that act-functions represent patterns of outcomes *directly*.[67] Some of these patterns may correspond to things that an agent might actually *do*, and some might correspond to things she might imagine herself doing, but many may not. It seems plausible to suppose, as Shafer suggests, that arbitrary *patterns of outcomes* are in principle available to the imagination, and that we might have preferences over such things, regardless of whether we can imagine any acts which might bring such patterns about.

One way to cash this idea out in more detail would be to let each act-function stand for an immense (possibly infinite) conjunction of counterfactuals,

$$(s_1 \mathbin{\Box\!\!\rightarrow} o_i) \;\&\; \ldots \;\&\; (s_n \mathbin{\Box\!\!\rightarrow} o_j)$$

It is then to be supposed that ‘$(s_i \mathbin{\Box\!\!\rightarrow} o_i)$’ is one of the conjuncts just in case the conjunction it forms a part of corresponds to the act-function which maps s_i to o_i.[68] It could then be said that:

> $(\{s_i\}, o_i \,|\, \ldots \,|\, \{s_n\}, o_n) \succcurlyeq (\{s_j\}, o_j \,|\, \ldots \,|\, \{s_m\}, o_m)$ if and only if the S prefers that $(s_1 \mathbin{\Box\!\!\rightarrow} o_i)$ & … & $(s_n \mathbin{\Box\!\!\rightarrow} o_n)$ rather than that $(s_1 \mathbin{\Box\!\!\rightarrow} o_j)$ & … & $(s_n \mathbin{\Box\!\!\rightarrow} o_m)$

Patterns of outcomes are not the kind of things that an agent *does*, nor are they the immediate objects of choice in any practical sense—so, again, this way of interpreting the elements of $\mathcal{A}$ does not sit well with a behavioural interpretation of $\succcurlyeq$. Note, however, that on this interpretation of act-functions there can be no question as to whether **SAV0** is true: every act-function can be uniquely paired with some conjunction of counterfactuals, regardless of what the decision-maker’s situation happens to be like.

There are complaints that can be raised, though. As Joyce (1999, 107-8) notes, it’s exceedingly unlikely that anyone’s preferences understood as such would satisfy **SAV1**, which requires $\succcurlyeq$ to be *complete* on $\mathcal{A}$. For one thing, there are far too many patterns of outcomes to imagine—uncountably many in Savage’s system, as it turns out—and there seems to be no rational reason to consider all of them. This point has both descriptive and normative force. Joyce argues that completeness is not a requirement of rationality, but

[67] In §5.4, I provide an argument from another direction that the best interpretation of Savage’s act-functions is in terms of patterns of possible outcomes which may or may not correspond to things the agent in question might do.

[68] Joyce (1999, 62-5) argues that counterfactual conditionals would be inadequate for this way of interpreting Savage’s act-functions, and instead posits a (somewhat mythical) ‘Savage conditional’ to play the role instead. It is orthogonal to my purposes to consider whether his argument against the use of counterfactuals is convincing, as the point I wish to make can be made just as well if we assume that every act-function represents an immense conjunction of Savage conditional statements.

it's all the more clear that completeness is not even close to descriptively plausible either—and this places pressure on any version of characterisational representationism based on Savage's theorem (or a theorem which requires a similarly rich space of act-functions). Furthermore, without completeness, it's unclear whether agents would non-trivially satisfy Savage's other preference conditions. Note that almost *every* Savage-like theorem assumes **SAV1**; indeed it's very difficult to achieve a strong representation result without it. Those that try to do without **SAV1** appeal to a notion of coherent extendibility (discussed shortly) and have correspondingly weak uniqueness results; see, e.g., (Seidenfeld, Schervish *et al.* 1990, 1995).

Indeed, there is a tension within Savage's system, between requiring that agents have complete preferences on the space of *imaginable acts* (or *imaginable patterns of outcomes*) on the one hand, and how their decision-making behaviour is modelled on the other. The set of null events $\mathcal{N}$ is intended to characterise those propositions that the agent has no credence in, and a decision-maker who satisfies Savage's preference conditions is modelled as essentially *ignoring* null events when choosing between her options—hence she is indifferent between two act-functions if their outcomes only differ on null events (Definition 5.10). Introspectively, this is plausible—when deciding between options we discount the impossible (and perhaps even the exceedingly unlikely). It is odd, then, to simultaneously require of an agent a disposition to discount zero credence states when considering an acts' outcomes, while at the same time require interesting preference patterns between acts she is sure she cannot perform (or patterns of outcomes she is sure cannot be brought about).

By dropping Act–Function Correspondence and reconstruing the interpretation of $\mathcal{A}$ as either a space of imaginable acts or arbitrary patterns of outcomes, all that has been achieved is the exchange of one problem for a host of others. While **SAV0**/**SAV0'** seems salvageable under the re-interpretation, it comes at the cost of making **SAV1** almost certainly false, and doubt can be cast on whether the remaining preference conditions can be non-trivially satisfied. There is, however, one further response to the constant acts problem which I will consider briefly, which seems to me the strongest response available to the proponent of characterisational representationism.

5.2.4 Coherent extendibility

Suppose that $\succcurlyeq$ is incomplete on $\mathcal{A}$, however $\succcurlyeq$ and $\mathcal{A}$ are supposed to be interpreted. This may be because $\succcurlyeq$ is given a behavioural interpretation and can only be coherently understood as holding between act-functions which correspond to available acts, and so is not defined on act-functions which don't correspond to available acts. (That is, $\mathcal{A}$ might be taken to represent the union of $\mathcal{A}'$ with some set $\mathcal{A}^*$ of purely fictional entities, where any behavioural preference relation would be defined only for pairs taken from the subset $\mathcal{A}'$.) Alternatively, we may suppose that $\succcurlyeq$ is incomplete on $\mathcal{A}$ because $\succcurlyeq$ is defined in terms of preferences between patterns of outcomes, but the agent only has preferences for a limited number of such patterns.

In any case, if $\succcurlyeq$ is incomplete on $\mathcal{A}$ then **SAV1** is false, then many of Savage's other preference conditions may be only trivially satisfied, and $\succcurlyeq$ is likely too impoverished to guarantee that $\succcurlyeq^{\mathrm{u}}$ and $\succcurlyeq^{\mathrm{b}}$ are complete on $\mathcal{O}$ and $\mathcal{E}$ respectively. Nevertheless, there may be an *extension* of $\succcurlyeq$, call it $\succcurlyeq^{+}$, which *does* satisfy all of Savage's conditions. Define an extension $\succcurlyeq^{+}$ of $\succcurlyeq$ as any superset of $\succcurlyeq$; thus $\succcurlyeq^{+}$ agrees with $\succcurlyeq$ regarding all those elements of $\mathcal{A}$ for which $\succcurlyeq$ is defined. If any extension of $\succcurlyeq$ conforms to Savage's conditions, then Theorem 5.1 entails that it can be given an expected utility T-representation. This fact could prove useful for characterisational representationism in dealing with the issues raised in §§ 5.2.2–3.

Say that $\succcurlyeq$ is *coherently extendible* if it has at least one extension $\succcurlyeq^{+}$ which *does* satisfy Savage's conditions (or the preference conditions of whatever theorem we are considering). It is not at all obvious that the preferences (however understood) of ordinary agents *are* coherently extendible with respect any contemporary Savage-like theorem's preference conditions—but if they are, then the path is open for the advocate of characterisational representationism to attempt a characterisation of credences and utilities in terms of the representations that the theorem supplies for the extended relations $\succcurlyeq^{+}$.

In most cases, if $\succcurlyeq$ is coherently extendible at all, then there will be a large number of extensions which satisfy the stated conditions, and something would have to be said about this fact. As suggested in Chapter 4, however, non-uniqueness is not a fundamental problem for characterisational representationism—so long as a theorem gives us substantial restrictions on the range of available interpretations, it need not have the Standard Uniqueness Condition. One could appeal to further information to filter between alternative extensions of an agent's $\succcurlyeq$, thus (assuming the theorem in question has strong uniqueness results) arriving at a single expected utility representation of the agent's preferences. Alternatively, it could be argued that agents' credences (and likewise their utilities, *mutatis mutandis*) are best represented by a *set* of probability functions—*viz.*, the set determined by each coherent extension of her preference relation. This idea is not new; in the literature a set of probability functions designed to represent an agent's total credence state is called her *representor*. For discussion, see (Levi 1974), (Williams 1976), (Jeffrey 1983), (Walley 1991), and (van Fraassen 1990, 1995).

Appealing to coherent extensions of $\succcurlyeq$ seems to me the best hope that characterisational representationism has for dealing with potentially incomplete preference systems—both for theorems within the Savage paradigm, and other theorems besides. For the strategy to be successful, of course, the preferences of ordinary agents must be coherently extendible to begin with—but it hardly seems like an impossible task to construct preference conditions such that this is possible. Moreover, to avoid a conflict with desideratum **(3)**, $\succcurlyeq$ will have to be defined on *enough* of $\mathcal{A}$, however it is interpreted, that the range of possible coherent extensions is substantially restricted. This may not be so, for instance, if act-functions correspond to *infinite* conjunctions of counterfactuals, as in §5.2.3—in which case, an ordinary agent may have *no* preferences over $\mathcal{A}$, so *every* way of satisfying the relevant preference conditions will be a coherent extension of her $\succcurlyeq$-ranking, and the representor will be utterly uninteresting *qua* model of her credences and utilities. Similar things are likely to be true if $\succcurlyeq$ is defined on the union of $\mathcal{A}'$ with some set $\mathcal{A}^*$ of purely fictional entities, *if* $\mathcal{A}^*$ constitutes the very large majority of $\succcurlyeq$'s domain.

Let me summarise where things stand with the constant acts problem in relation to characterisational representationism. Admitting imaginary act-functions into $\mathcal{A}$ and assuming Act–Function Correspondence is not a coherent possibility. Thus, the proponent of characterisational representationism might try to retain Act–Function Correspondence while reconstructing $\mathcal{A}$ from the ground up *à la* Lewis, or she might drop Act–Function Correspondence and supply some alternative interpretation of $\mathcal{A}$ and $\succcurlyeq$.

Either option is consistent with appealing to a notion of coherent extendibility to achieve a final representation of an agent's credences and utilities. Appealing to coherent extensions will in general mean giving up on using the theorem to construct a unique $\mathcal{B}el$ and $\mathcal{D}es$ model of the agent, but given the kinds of strong preference conditions needed to attain strong uniqueness results that was likely a fool's errand in any case. The appeal to coherent extensions also suggests the possibility of retaining a behavioural interpretation of $\succcurlyeq$ even without Act–Function Correspondence. It is less clear, however, whether ordinary agents' preferences over whatever $\mathcal{A}$ represents are (a) coherently extendible to begin with, and (b) sufficiently rich so as to substantially narrow down the range of possible coherent extensions.

5.3 States, events, and the objects of credence

The next problem that I will discuss concerns the domain of $\mathcal{B}el$. In this section, I will first outline the problem as it arises within Savage's theorem in particular, before generalising to other theorems in the broader Savage paradigm. I will argue that these theorems do not allow us to assign credence values to enough propositions, or to the right kinds of propositions, to adequately represent anybody's total credence states: credences are only assigned to *disjunctions of states*, and many of the most interesting propositions—including those about acts and outcomes—cannot be expressed as a disjunction of states.

That the $\mathcal{B}el$ function derived using Savage's theorem in particular does not supply credence values for acts has been noted before (see for example Spohn 1977, 117-8, Joyce 1999, 117); indeed those who accept the 'crowding out' thesis that I will discuss in §5.3.2 sometimes see this as a unique advantage that Savage's decision theory has over others. This characteristic of Savage's theorem is usually attributed to his assumption that states are act-independent. As I will show, however, the same property attaches to all (single-primitive) representation theorems which make use of the same basic formal structures that Savage employs—even those which don't assume that states must be act-independent in any ordinary sense. The problem here is not unique to Savage, and cannot be avoided just by tweaking some of his background assumptions.

5.3.1 The domain of Savage's $\mathcal{B}el$

I will assume, for now, that both **SAV0**/**SAV0'** and Act–Function Correspondence are true. I will argue shortly that such strong assumptions aren't needed to bring out the issue her discussed, but it's easiest to begin with them nonetheless. This implies that states are outcome-independent, for otherwise Savage's act-functions don't make sense *qua* representations of possible acts an agent might take. Likewise, I will set aside concerns about whether ideal or non-ideal agents satisfy Savage's preference conditions—I will assume that everyone does. And finally, to avoid the worry that ordinary agents' total credence states might not be representable by probability functions, I will even assume that non-ideal agents tend to be probabilistically coherent. These assumptions load the dice very much in favour of an appeal to Savage's theorem as a basis for characterisational representationism. However, even if they are granted, there is a further problem: the domain of $\mathcal{B}el$ is simply not rich enough to allow for the representation of our full range of credence states.

In Savage's system, $\mathcal{B}el$ is defined only for *events*. Every event corresponds directly to a particular proposition—in particular, to some disjunction of states—but, crucially, not every proposition corresponds to an event. In what follows, I will refer to propositions which don't correspond to an event as *non-event propositions*. The question is whether these non-event propositions form an important class, with members towards which the ordinary subject does (or can) have credences. There are two kinds of propositions to consider; namely, those regarding what acts she might perform, and those regarding the outcomes that might result.

That Savage's $\mathcal{C}r$ is undefined for propositions regarding what acts we might perform follows immediately from the assumption of act-independence. Since the choice of (and performance of) any one act α implies foregoing the other options on the table, every state is consistent with the performance and non-performance of α. Thus, the line that divides *α is performed* from its negation cuts across the lines that divide states from one another—neither proposition is equivalent to any disjunction of states. But—as I will argue in more detail below—these kinds of propositions certainly do seem like things that we can have opinions about!

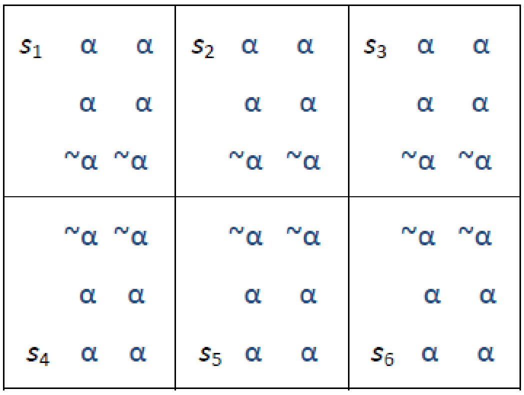

Figure 5.1

The above toy model will help to bring out this point (Figure 5.1). The entire rectangle is the set of all possible worlds, partitioned into a number of states ($s_1 - s_6$), each of which contains six different worlds—some are worlds where *α is performed* (represented by α), and some are worlds where *α is not performed* (represented by ~α). Every event corresponds to some collection or other of states; for instance, {s_1}, {s_3, s_4}, or {s_1, s_3, s_5}. However, the proposition that *α is performed*—the set of all α worlds—does not correspond to any collection of states: it is a non-event proposition. We can make ever finer distinctions between states, but as long as every state has both worlds where *α is performed* and worlds where *α is not performed*, neither proposition will ever correspond to any event. States are just not fine-grained enough to make the relevant distinctions between possibilities.

The same can be said for outcome-propositions. Each outcome $o \in \mathcal{O}$ is distinct, and if one outcome obtains then no other outcome does. By the same reasoning that we have just seen, then, states don't cut finely enough to make the relevant distinctions we need here. The same applies to *any* proposition I care about, the truth of which is at least partially dependent on my choices. For instance, suppose that some outcomes are *nice*, while other outcomes are *nasty*. Then, the proposition *something nice happens* is a non-event proposition: every state is consistent with nice things happening and also with nasty things happening, so there is no way to form that proposition as a disjunction of states. Or, perhaps I care about whether I get to eat tomorrow, and this is not guaranteed to occur independently of my actions. Then *s* will be compatible with both *I will eat dinner tomorrow* and *I will not eat dinner tomorrow*, so neither proposition is an event—although I certainly *do* have credences (high credences, in fact) that I will be eat dinner tomorrow.

It might be supposed that $\mathcal{B}el$ does manage to supply an accurate representation of the preference-rational subject's credences with respect to *events* only, despite falling silent regarding non-event propositions. This is problematic, given just how many important non-event propositions there are—to recall, outcomes may make reference to "money, life, state of health, approval of friends, well-being of others, the will of God, or anything at all about which the person could possibly be concerned". Taking this line would mean that by using Savage's theorem, one could at most only arrive at a very partial characterisation of what credences are—and many of the most interesting credence states will need to be characterised in some other way. This kind of retreat to a merely partial characterisation seems poorly motivated, though. In particular, there seems to be no good reason to think that the metaphysics of credences should be disjunctive, in the sense of involving one account for what it is to have a credence of x towards one class of propositions and another account for what it is to have a credence of x towards the rest. If the representation theorem does not give us enough to characterise *all* of the credence states that we have, then it seems rather more sensible that we should seek some other, more general account of the nature of credences.

Indeed, it is unclear whether ordinary agents have credences with respect to many events at all. On Savage's conception, states are very odd creatures—and so too, therefore, are most events.[69] Recall, for instance, the characterisation of states as dependency hypotheses (§5.2.2): to satisfy Savage's required conditions, individual states must have something like the character of a dependency hypothesis, yet it's doubtful that any ordinary agent is able to contemplate even a single dependency hypothesis let alone have credences regarding them. The point is all the more convincing for arbitrary disjunctions of dependency hypotheses. Of course, the ordinary agent will likely have credences for the necessary event, $\mathcal{S}$, and the impossible event, $\emptyset$; but it's doubtful that she will have credences for more complicated events. Note, of course, that Savage's $\mathcal{B}el$ function is defined for *all* events—so besides being impoverished in one sense, the domain of Savage's $\mathcal{B}el$ also seems *too rich* in another.

So much for the problem as it arises in Savage's system. More than one implausible assumption went into the above argument; perhaps the right lesson to draw is that one would do best to not appeal to Savage's theorem when developing characterisational representationism—something that should already be obvious given fact that Savage's $\mathcal{B}el$ is limited to probability functions. However, the problem just outlined goes beyond Savage's theorem. To see this, note that to raise the central worry here we don't need to assume that *every* state is compatible with *every* act, nor that *every* state is compatible with *every* outcome. If so much as *one* state is compatible with both P and $\neg P$, then P and $\neg P$ are non-event propositions. It would be enough to make the point to simply establish that *there are* states compatible with multiple, mutually inconsistent acts and outcomes.

For example, in the following model, although only one state (s_1) is consistent with both *α is performed* and its negation, the proposition *α is performed* (the set of α worlds) is not equivalent to any disjunction of the states:

[69] This is a point which has long been known; see, for instance, (Balch and Fishburn 1974, 57-8) and (Joyce 1999, 118).

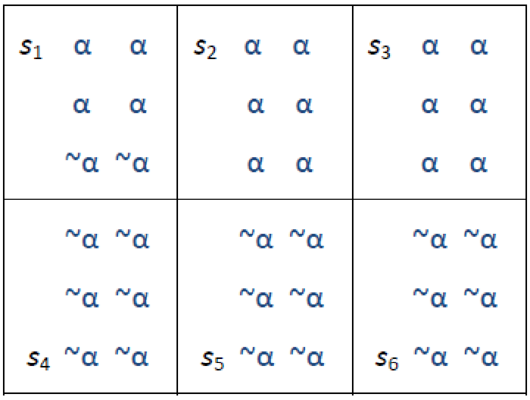

Figure 5.2

For the problem to arise, all that is needed is that one or both of the following two conditions are satisfied:

(A) There are *acts*, about which *S* has credences, such that at least one state exists that is consistent with the performance and the non-performance of that act

(B) There are *outcomes*, about which *S* has credences, such that at least one state exists that is consistent with that outcome obtaining and it not obtaining

Neither (A) nor (B) imply that every state is consistent with every outcome, nor even with every act. Their satisfaction is compatible, for example, with supposing that every outcome is a maximally specific proposition (or just an act-state conjunction), such that every outcome is consistent with exactly one state.

These are very weak conditions, and their presence in any Savage-like representation theorem can be assumed for very good reasons. The motivation for (B) is obvious. The point of decision theory applied to situations of uncertainty is to determine which choice to make on the basis of the different outcomes that each available act would have, given each of the different states that are consistent with what we know to be true. A quick glance at the standard decision matrix (§2.4) will reveal that the framework is useless if every act has exactly the same outcome at a state as any other act. Dominance reasoning, for example, would be impossible, as no act could do better at a state than any other. Likewise, if (A) were false then there would be no sense in applying decision theory in the first place—each state would *determine* that a particular choice was made, so there would be no meaningful comparison of the outcomes of different acts at a state.

More generally, we ensure that states are consistent with multiple options and with multiple outcomes because it is a basic presupposition of decision theory that we are able to freely make choices between alternative acts with interestingly different consequences dependent upon the true state of the world, of which we are uncertain. But where the true state (whatever it may be) *entails* that a particular choice was made, and that any alternative choice would have resulted in the same outcome anyway, then there was never the possibility to choose otherwise or even a reason to contemplate the choice in the first place. There would be no point in quantifying our uncertainty about states if neither of (A) and (B) were true, as the true state would not afford us a choice. Decision-making is incompatible with this kind of fatalism. When making a decision regarding what act to perform, we engage in a minor fiction: that the actual state of the world is compatible with multiple acts being chosen, each with potentially different outcomes.

The extent of the problem obviously depends on *how many* acts and outcomes satisfy the two conditions. If there were only *one* act which made (A) true, and *one* outcome which made (B) true, then the problem would not be very widespread—the subject could not be represented as having credences about *that* act, or about *that* particular outcome, but $\mathcal{B}el$ would be capable of representing credences about other acts and outcomes. A bullet worth biting, perhaps. But the foregoing reasons indicate that the problem is extensive, not limited to one or a few acts and outcomes, but applicable to most acts and outcomes *at least*. Some special acts might be logically entailed or logically inconsistent with some states, and likewise for some special outcomes, but a decision-theoretic framework will not be widely applicable if this is true of most acts and outcomes.

For a Savage-like theorem to avoid the worry being raised here, it would need to be the case that each state implies that a particular act was chosen, and likewise that a particular outcome obtains—and for that matter, each state would have to imply either P or $\neg P$, for any proposition P that we suppose the agent can have credences about. No such theorem presently exists, and it's difficult to imagine what one would look like; at the very least, the familiar Savage-style representation of acts as functions (or partial functions) from $\mathcal{S}$ to $\mathcal{O}$ would be off the table. The majority of theorems—for both expected and non-expected utility theories—closely follow Savage in this way of characterising the basic objects of preference (see, for example, Stigum 1972, Gilboa 1987, Schmeidler 1989, Tversky and Kahneman 1992, Buchak 2013, Alon and Schmeidler 2014). Some of these theorems manage to avoid appealing to constant act-functions, but all imply both (A) and (B). Luce and Krantz's (1971) theorem departs slightly from Savage's paradigm by representing acts as *partial* functions from $\mathcal{S}$ to $\mathcal{O}$, but while they explicitly reject Savage's act-independence assumption, their states are still consistent with multiple acts and outcomes.[70] Despite their differences, though, even in these systems both (A) and (B) are implied and the problem raised here applies.

In his (2014), Kenny Easwaran adapts the Savage paradigm and in effect represents acts as partial functions from $\mathcal{S}$ to $\mathcal{O}$, requiring in particular that distinct act-functions are either defined on precisely the same, or wholly disjoint, subsets of $\mathcal{S}$. It seems consistent with his results to have each act-function defined on a distinct subset of $\mathcal{S}$, thus falsifying condition (A) and thereby also (B). However, Easwaran is able to derive interesting results from preferences between his act-functions only because he assumes the existence

[70] This is a direct consequence of their act-richness assumptions, specifically axiom 1 of definition 1.

of a primitive $\succcurlyeq^b$ relation. Through $\succcurlyeq^b$, Easwaran sets up "correspondences" between disjoint events via their similar credence values, thus allowing for fruitful comparisons between act-functions defined for wholly distinct sets of states. Because of this, Easwaran's construction is unsuitable as a basis for characterisational representationism (a fact that he is explicitly content with). It is doubtful that any interesting representation theorem could be developed using this kind of formal structure without appealing to something like this correspondence relation.

5.3.2 Deliberation and prediction

Perhaps the problem is not as bad as I have made out—there are, after all, some who argue that we lack credences regarding whether we will perform one or another of the acts currently available to us in a given choice situation. Wolfgang Spohn, for example, claims that "probably anyone will find it absurd to assume that someone has subjective probabilities for things which are under his control and which he can actualise as he pleases" (1977, 115). Spohn's claim is that because it is entirely under her control whether *S* chooses to perform a given act or not, there is no sense in her being *uncertain*—or *certain*—about whether the act will be enacted; she simply lacks those credence states. Let us refer to this as the *Deliberation Crowds Out Prediction* (DCOP) thesis; besides Spohn, it has also been advocated by (Levi 1989, 2000, 2007), (Gilboa 1994), and more recently, (Price 2012), and (Ahmed 2014).

To continue the thought, outcomes might also be conceived of as being closely connected to acts, in such a way that if we were to lack credences in the latter then we might plausibly lack credences in the former. Indeed, Spohn (1977, 116) argues for precisely this. His argument presupposes that agents' credences can be represented by a probabilistically coherent credence function, $\mathcal{P}r$, such that for any pair of propositions P and Q in $\mathcal{P}r$'s domain,

$$\mathcal{P}r(P) = \mathcal{P}r(Q).\mathcal{P}r(P|Q) + (1 - \mathcal{P}r(Q)).\mathcal{P}r(P|\neg Q)$$

If this were true, then if the agent had credences regarding some proposition P which probabilistically depends on her performance of an act α, she would be able to indirectly induce an unconditional probability regarding α using the above equality; hence, if she does not have credences regarding α, she cannot have credences for any such P. Of course, the generality of this argument is questionable: ordinary agents are not plausibly probabilistically coherent, and it's not clear whether even ideally rational agents ought to be either.

In any case, though, the important point is that there may be ways to tie credences about outcomes to credences about acts in such a way that a lack of credences with respect to the latter plausibly leads to a lack of credences with respect to the former. (For example, if outcomes were simply act-state conjunctions, then plausibly there should be no credences for outcomes inasmuch as there are no credences for acts.) If so, then the truth of DCOP would certainly diminish the force of the problem raised in §5.3.1. Indeed, the fact that Savage's $\mathcal{Bel}$ will not represent credences about such things would be a particularly attractive *feature* of applying his framework—the relevant credences states never existed to begin with!

I do not share Spohn's sense of absurdity that is supposed to come with ascribing credences to S regarding acts that are presently under S's complete control to perform, should she so choose. One of the strongest arguments (read: not based on the betting interpretation) for the DCOP thesis seems to be that credences regarding which action will be chosen in the present circumstances *play no role* in rational decision-making and so there is no theoretical reason to posit such states (Spohn 1977, 114-5). Even supposing that this is true—it may be in *Savage*'s decision theory, but of course there are alternatives (e.g. Jeffrey); see also Rabinowicz (2002, 112-4) and Joyce (2002) for a critique of this claim—it is one thing to say that credences about acts play no role (or *should* play no role) in decision-making and quite another to say that we simply *don't have* such credences.

By way of example, note that utilities for *events* also play no role in decision-making according to Savage's decision theory. By hypothesis, what event obtains is independent of the choice between acts, so any valuation of the events on the subject's behalf is irrelevant to her choice. It would be unreasonable to infer from this that we *don't have* utilities for events; at least, it certainly seems to me that I am able to judge which of two events I would rather be true, even if I know that this is entirely beyond my control. One of the central theoretical roles of utility assignments is to represent a subject's preferences over ways the world might be—that such states may not play a role in rational decision-making (according to Savage) does not mean that there is no reason to posit them at all. Likewise, I seem to be able to ascribe credences about my own actions to myself, even during deliberation. On the basis of past evidence, I know that when I am faced with the decision between caffeinated and non-caffeinated beverages, I tend to choose the former; surely, were I in that situation now, I could be *confident* that I would do the same—and I should be able to represent such confidence in my credence function. I may even surprise myself with an herbal tea on occasion!

Indeed, denying the existence of these credence states comes with severe theoretical costs. For instance, the thesis is in conflict with the principle of Conditionalisation. The actions that we might make in *future* situations aren't under our complete control *now*, and neither are the actions that we have *already* made. So we can have credences with respect to future and past actions. This is as it should be—in many circumstances, we ought to take credences about our past and future actions into account when making decisions. It is only credences about the actions that we might *now* perform which are ruled out by the thesis that deliberation crowds out prediction, as it's only those which are completely under our present control. But *this* certainly seems odd: I am confident now that I shall choose the caffeinated beverage when the option is available tomorrow; and tomorrow, after I have chosen the caffeinated beverage (probably), I shall be confident of having done so—but for that brief moment when I make the choice, my credences regarding that act will vanish from existence, only to reappear a moment later. Conditionalisation will not explain such changes. For similar reasons, if we necessarily lack credences regarding acts (and outcomes!) then we are only a short step away from counterexamples to both the General Reflection Principle (van Fraassen 1995) and the Principal Principle (Lewis 1980b).[71,72]

There may be some sense in which 'deliberation crowds out self-prediction', but whatever that sense may be, it's *not* the sense in which we simply lack credences about acts. Rabinowicz (2002, 92-3), for example, suggests that perhaps credences about acts "are available to a person in his purely cognitive or doxastic capacity, but not in his capacity of an agent or practical deliberator"; that is, while the agent *does* have credences about acts, *while* deliberating about what to do these credence states are (for whatever reason) cognitively inaccessible. This may be more plausible if we distinguish between *explicit* and *conscious* assignments of credence values to acts about propositions from what we might call *standing* or *implicit* credence states. Alternatively, one might try to establish that *if* an agent *S* has credences regarding acts, then she ought not to *consider* or *use* those credences whilst deliberating—*rational deliberation* crowds out the *consideration* and/or *application* of certain credence states, perhaps, but not their existence.

Moreover, whatever we might say about credences towards acts, there is still the problem with outcome propositions. To recycle the earlier examples, I may or may not eat dinner tomorrow, depending on what I choose to do now. If any state at all is compatible with both of these outcomes, then *I will eat dinner tomorrow* is not an event, and it will not be in $\mathcal{B}el$'s domain.[73] Whatever might be said about credences in acts specifically, we certainly *do* have credences regarding non-event propositions.

[71] Advocates of DCOP usually allow that we can have credences conditional on propositions about what acts we might perform; e.g., *S* should have a credence of 1 that she will perform α given that she performs α. This makes DCOP incompatible with the orthodox definition of conditional probabilities, but this is hardly a severe cost—as Hájek (2003) shows, there are strong independent reasons for rejecting the orthodox definition.

[72] Because it is already incompatible with conditionalisation, accepting DCOP will not bring us into conflict with Rachael Briggs' (2009) Qualified Reflection principle, which states that "an agent should obey Reflection only if she is certain that she will conditionalise on veridical evidence in the future" (59). This is the right result, but for entirely the wrong reason.

[73] A further point is relevant here: That I perform a particular act in the future is not a guaranteed outcome of any present actions I may now take, so the problem extends even to credences about future acts. This is obviously problematic for the reason mentioned above: Many decision situations seem to require taking such credences into account. Thanks to Alan Hájek for pointing this out.

To sum up: both $\mathcal{B}el$ and $\mathcal{D}es$ appear to have inadequate domains if they are to serve as representations of ordinary agents' credences and utilities respectively. The origin of the problem, of course, is the representation of acts as functions from $\mathcal{S}$ to $\mathcal{O}$. This aspect of the Savage paradigm accounts for much of its popularity: act-functions provide a purportedly straightforward means of connecting acts to objects of uncertainty (states and events) and objects of utility (outcomes), all the while allowing theorists to characterise behavioural preferences over acts in a manner that *appears* to make the relation transparent to empirical observation. In the next section I will argue that such appearances are misleading, but for now the important point is that the use of act-functions comes with a cost. If these functions are to represent acts, then constraints must be placed on $\mathcal{S}$ (and hence $\mathcal{E}$) and $\mathcal{O}$—constraints which are ultimately manifest in the limitations of the $\mathcal{B}el$ and $\mathcal{D}es$ functions derived from $\succcurlyeq$ on $\mathcal{A}$.

5.4 Acts and intentionality

The proponent of characterisational representationism wants to be able to say that their chosen representation theorem allows us to characterise what it is for an agent to have such-and-such credences and utilities by reference to her preferences. With this goal in mind, then on pain of circularity the interpretation of the theorem's basic elements had better not involve some (tacit or explicit) appeal to the agent's credences or utilities. It should be possible, that is, to understand and specify the basic notions involved in the statement of the theorem without any prior knowledge regarding her credences and/or utilities. This was our desideratum **(4)**, outlined in §3.4.5. Many proponents of characterisational representationism will also want to say that the relevant preferences are *behavioural*, and that something like Savage's framework lends itself well to the project of naturalising away the mental—that is, that Savage-like theorems satisfy desideratum **(5)**.

I will argue that neither of these desiderata will be met by any Savage-like representation theorem—at least inasmuch as desideratum **(2)** has a hope of being satisfied.[74] This is contrary to first appearances, as the basic framework that Savage described seems to be well-suited for a fully naturalistic interpretation—indeed it was designed with a behaviouristic definition of credences and utilities specifically in mind, and theorems within the paradigm are still today treated as underwriting behavioural definitions of credence and utilities. We take a subject, *S*, in a given decision-making context. From a purely physical standpoint we describe—with an appropriate degree of specificity—the set of acts ($\mathcal{A}'$) that *S* might perform in that context, such that each one precludes the performance of any other in the set. $\mathcal{A}'$ forms the domain of a behaviourally-interpreted $\succcurlyeq$. We

[74] In fact, I will argue that desiderata **(2)**, **(4)** and **(5)** cannot be simultaneously met by *any* current representation theorem, for essentially the same reasons discussed in this section. The best that characterisational representationism can do is try to satisfy **(2)** and **(4)**, given how things currently stand. See Chapter 9.

can observe which of these acts S *actually* decides to perform, and this will sit at the top of her $\succcurlyeq$-ranking. This gives us some information, but not enough. To get at the rest of her preferences, we suppose that S's choice-dispositions regarding pairs of acts in $\mathcal{A}'$ can be determined without reference to her mental states, thus supposedly giving us everything we need to construct an effectively unique $\mathcal{B}el$ and $\mathcal{D}es$ under the assumption that S maximises expected utility.

That is a common conception of how a Savage-like theorem is applied towards the behavioural characterisation of credences and utilities. Constant act-functions and other imaginary act-functions are treated as a convenient fiction, to be dealt with perhaps in some future patched-up theorem or explained away by reference to coherent extensions—but either way they aren't taken to be especially devastating for the intended behaviouristic interpretation of $\succcurlyeq$ and its relata. For the most part, the received wisdom is that Savage has *basically* shown us how to work backwards from a subject's behavioural dispositions to a unique representation of her credences and utilities.

But things are not so simple. The acts as they appear in $\mathcal{A}'$ are things like *walk to work*, *drive to work*, *skip to work*, *stay at home*, and so on. However, Savage's framework requires $\succcurlyeq$ to be defined over entities with a particular formal structure—i.e., act-functions—so in order to make use of any Savage-like theorem we first need to associate each act in $\mathcal{A}'$ with a unique act-function—and therein lies the origin of the problems to be discussed in this section. The central issue of this section concerns the *right* way associate act-functions with acts.

There are two options for associating act-functions with acts that I will discuss. First, each act-function $\mathcal{F}_\alpha$ might represent an act α by picking out α's *actual causal profile*. Second, $\mathcal{F}_\alpha$ might represent α by picking out α's causal profile *as the decision-maker understands it to be*. Neither option is forced upon us by Savage's formalism, so the choice depends on which will afford the more useful interpretation of his theorem. In §5.4.2, I will argue that only the second option is viable, if desideratum (2) is to be satisfied. However, in what follows, I will argue that regardless of which option we pick, there is no way to arrive at an adequate association of act-functions and acts without some prior access to certain of the subject's intentional mental states—including, most importantly, some of her doxastic states.

5.4.1 Specifying states and specifying outcomes

In §5.1.1, I outlined the standard way in which act-functions are assumed to be associated with acts. Given first an appropriate specification of $\mathcal{S}$ and $\mathcal{O}$, each act α in an appropriately specified set $\mathcal{A}'$ can be associated with a unique act-function which essentially formalises α as *the act which would lead to such-and-such outcomes if it were performed in such-and-such states*. That is, given a choice of $\mathcal{S}$ and $\mathcal{O}$, we can always pair each α in $\mathcal{A}'$

with a unique act-function that represents α's *actual causal profile* if the following two conditions hold:[75]

(i) All s in $\mathcal{S}$ are *logically independent* of the performance of the acts in $\mathcal{A}'$

(ii) All s in $\mathcal{S}$ are *outcome-functional* with respect to the specification of the acts in $\mathcal{A}'$ and the outcomes in $\mathcal{O}$

These two constraints are not *necessary* for being able to map acts to unique act-functions, but they are sufficient—so long as (i) and (ii) hold, there will be an act-function which uniquely corresponds to any act α in $\mathcal{A}'$. The key point for what follows, however, is that given (i) and (ii), the appropriate specification of $\mathcal{S}$ is constrained by the specifications given of $\mathcal{A}'$ and $\mathcal{O}$. The same holds true for any Savage-like representation theorem given the minimal conditions (A) and (B) noted at the end of §5.3.1. For this reason I will simplify the discussion and assume that (i) and (ii) hold in general.

Importantly, while it's true that given (i) and (ii) there will be only one act-function *specifically from* $\mathcal{S}$ *to* $\mathcal{O}$ which corresponds to a given α in $\mathcal{A}'$, there may be many other act-functions defined for *different* sets $\mathcal{S}^*$ and $\mathcal{O}^*$ which *also* directly correspond to α—and there is more than one possibility for what $\mathcal{S}^*$ and $\mathcal{O}^*$ could be (Levi 2000). There are also different ways of carving up the range of acts $\mathcal{A}'$, depending for instance on the degree of *specificity* with which each act is described. The $\mathcal{B}el$ and $\mathcal{D}es$ pair that we end up with depends heavily on the particular way in which $\mathcal{S}$ and $\mathcal{O}$ are specified, so much depends on getting it right. I will argue that there is no way to do this without having prior access to the agent's credences and utilities (or something nearby).

Let us first of all get the obvious problem out of the way: if $\mathcal{O}$ is a partition of propositions which are highly (if not maximally) specific with respect to what the agent *cares* about, then it is doubtful than any appropriate specification of $\mathcal{O}$ can be given without reference to the agent's utilities, desires, or (mentalistic) preferences. We need to assume that $\mathcal{O}$ tracks what the agent cares about, of course, because a basic presupposition behind any representation theorem is that agents choose between options following the consideration of those options' potential outcomes—so differences between outcomes which matter to the agent will matter to the final choice. (A similar problem arises, of course, for ensuring that outcomes are context neutral, but not every Savage-like theorem imposes this requirement.)

If we knew the decision-maker's *utilities*, it would be easy enough to work out which distinctions make a difference to how she values possible states of affairs. Likewise, if we had access to her *mentalistic* preferences between propositions, it would likely be

[75] I assume, like Savage, that two acts with exactly the same pattern of outcomes—if such a thing is even possible under a sufficiently fine-grained partition $\mathcal{S}$—can be treated as indistinguishable for the purposes of decision-making.

possible to work out which propositions are highly (or maximally) specific with respect to what she cares about. But to have either kind of information we would need access to her intentional states, and the whole *point* of appealing to behavioural preferences was to avoid reference to mentalistic preferences and presupposed utilities. Moreover, if we had access to her utilities or mentalistic preferences over *arbitrary* propositions, it would be a mystery why $\mathcal{D}es$ should be defined *only* on $\mathcal{O}$—and indeed, why this apparently very relevant information about her preferences should not be taken into account when constructing $\mathcal{B}el$ and $\mathcal{D}es$!

Note, furthermore, that in general it would not be plausible to treat the propositions in $\mathcal{O}$ as maximally specific *simpliciter*, and thus specific with respect to anything the agent might care about. Such a move would be in obvious tension with any theorem which, like Savage’s, has $\succcurlyeq$ defined on act-functions which pair outcomes with multiple states. (In §5.4.2, we will see another reason for thinking that the elements of $\mathcal{O}$ should not be *too* specific, if $\mathcal{D}es$ is supposed to capture the utilities of any ordinary agent.)

An equally worrying problem arises with the proper specification of $\mathcal{S}$. Besides being constrained by $\mathcal{O}$, any adequate specification of $\mathcal{S}$ is also constrained by $\mathcal{A}'$, and here we also see problems. As it was characterised in §5.1.1, $\mathcal{A}'$ ought to be a set of mutually exclusive acts which jointly exhaust the agent’s options in a given situation, described at a reasonably specific degree of detail, each of which—and this is the important part—is such that *the agent is certain that she would perform the act, if she were to intend as much*.

We will set aside, for now, issues relating to deciding *the right* degree of specificity when characterising the acts in $\mathcal{A}'$. Likewise, we will set aside any issues that might arise as a result of the reference to the agent’s counterfactual *intentions*—though this should itself be a cause of unease for characterisational representationism, especially if intentions are best understood in terms of desires and means-ends beliefs (on this view, see Anscombe 1963, Bratman 1987, Ridge 1998). Our concern regards the condition that, if α is to appear in $\mathcal{A}'$, then S should be *certain* of her capacity to perform α. If the set $\mathcal{A}'$ is restricted to acts which the agent in question has some sufficiently high degree of confidence in her capacity to perform, then we will need access to at least some of her doxastic states prior to the specification of $\mathcal{A}'$.[76]

[76] An alternative to the characterisation of acts I have given can be gleaned, with minor modifications, from Jeffrey (1968): An act α is *available* (i.e., in $\mathcal{A}'$) for an agent S just in case it would be rational for S to *become certain* that she has performed α by choosing to perform α—in the sense that her credence that she has performed α should be 1 after conditionalising on the evidence that she has chosen to perform α. This characterisation does not improve things greatly for characterisational representationism: The reference to what credences a rational agent would have under certain circumstances is still worrisome inasmuch as the goal is to understand and characterise what credence states are generally.

The motivation for this condition can be made evident with the help of an example:[77]

> Before Jill is a red button, above which is a sign reading 'PRESS ME FOR $100!' Jill knows that she can push the button easily, and also knows that the button will only do something if it's pushed—however she is not certain *what* it will do. As a matter of fact, the sign is accurate and pushing the button will cause $100 to pop up from a hidden compartment, free for her to take with no strings attached. Jill could do with the money, but she does not believe the sign: she knows that a prank-centred TV show is in town, and is (for good reason) rather more confident that she is on camera, and that pushing the button will only result in her receiving a painful electric shock or some other cruel outcome. Jill chooses to leave the button alone.

Clearly, Jill *would* have been able to press the button had she so intended, and she knows this. Furthermore, if she had so intended, she *would* have received $100 as a result of pressing the button. It would be admissible to let $\mathcal{A}$' be {*push the button*, *leave the button alone*}. But it would be problematic if we were to characterise $\mathcal{A}$' as {*receive $100 by pushing the button*, *leave the button alone*}. Jill needs the money, and if she *knew* that she could *receive $100 by pushing the button* then she most certainly would have chosen that option rather than preferring to leave the button alone. She did not push the button because she did not know that receiving $100 was one of her options.[78] If acts are characterised as those things which *S would* perform if she intended to, *such that S* is certain that she would be successful if she so chose, then *receive $100 by pushing the button* (and *receive painful electric shock by pushing the button*) will not be amongst Jill's available acts—but *push the button* and *leave the button alone* will be. This seems to be as things should be—otherwise it would be exceedingly odd that Jill's choices reveal a preference for not pushing the button over receiving $100.

To be sure, this restriction on what acts can go into $\mathcal{A}$' raises some interesting issues. For one thing, requiring that Jill is *certain* of her capacity to perform any act in $\mathcal{A}$' may rule out too much—there are very few acts which Jill is *absolutely* certain she can perform. Nevertheless, something like this restriction is required to make sense of the fact that in Savage's theorem (and all similar theorems), preference-rational agents are implicitly modelled as being certain of their capacity to perform any of the acts over which they have behavioural preferences: this is why the expected utility of performing an act α is calculated through consideration of α's—and *only* α's—potential outcomes. If the agent

[77] A similar case to this is Brian Hedden's 'Raging Creek' example, in (Hedden 2012, 347-8).

[78] If the reader is uncomfortable with treating *receives $100 by pushing the button* as an act, alternative examples which make essentially the same point are easy to come by. Ultimately, all that is required is a mismatch between the acts that are actually available to an agent and the acts she believes are available, where her choices would have been very different had she been aware of the facts regarding her available options.

gave some substantial credence to the thought that by *intending* to perform α, she might instead end up performing β, then presumably some consideration of β's possible outcomes should play a proportionate role in her deliberations about whether to try to do α.

Another problem case arises when *S* is certain that she can perform α, but—as a matter of fact—if she were to try she would fail. Perhaps Jill is mistakenly certain that pushing the button would destroy the universe. In this case, it's unclear whether *destroy the universe by pushing the button* should be included in Jill's range of available acts—Jill herself seems to think it is! To deal with such cases, Sobel (1986) has suggested that rational agents can never be certain of a falsehood, but regardless of whether that's true, the same is obviously false for the ordinary person on the street. A weaker suggestion would be that if rational agents are certain they can perform α, then they can. However, on any natural conception of an act, this still seems too strong—and it does not help us to characterise $\mathcal{A}$' for non-rational agents.

These and similar considerations lead Schwarz (2014c, 7-11) to suggest that decision theory is best thought of *not* as a theory about preferences over *acts*—conceived of as things like *go to the park*, *get a drink from the fridge*, and so on—but instead as a theory about preferences over *intentions*—specifically, intentions to act in different ways. Hedden (2012) defends a nearby view, though he casts his position in terms of 'decisions' rather than intentions. It seems somewhat more plausible, for example, to suggest that the ordinary agent has complete epistemic access to what *intentions* she might form, so that she can reliably be certain that if she decided to *intend* to perform α, then she would be successful in intending as such.

Applied to a Savage-style representation theorem, the idea would then be that act-functions are best understood as models of *intentions to act*. Intentions to act have variable causal consequences depending on the exact state of the world, so it is natural to think that functions from states to outcomes can be used to pick out intentions just as well as they pick out acts. However, intentions are also a kind of mental state, so if this is the right way to understand Savage's formalism then it puts the lie to any purportedly behaviouristic or otherwise fully naturalistic interpretation of his and other similar representation theorems.

In the next section, I will argue that act-functions should not be taken to *directly* represent either acts *or* intentions to act (or decisions to act)—these are things which an agent can *do*, but act-functions are better seen as models of how an agent represents the pattern of outcomes which result from the things she thinks she can do. We associate different patterns of outcomes with different things we might do, and on that basis make decisions, but the *doing* and our representation of its possible outcomes are quite separate phenomena. Every act-function can be associated with a unique act α, to be sure, but *only* by picking out α *as it is represented by the subject*.

For now, however, the upshot of the foregoing discussion is clear: regardless of whether act-functions are supposed to represent *acts* or *intentions* (as they in fact are), there does not appear to be any way of pinning down *the right* collection of act-functions—of specifying the right $\mathcal{S}$ and $\mathcal{O}$—without some prior access to what goes on inside the agent's head.

5.4.2 Acts, and our representations thereof

The issues raised in §5.4.1 are all ultimately consequences of the truism that in order to make sense of an agent's choices, it's important to take into account how the she *represents* the decision situation that she is in and the options available to her. Let us therefore generalise the point: act-functions should not be taken to represent acts *directly*, but acts as they are understood by the decision-maker. In what follows, I will argue that there does not appear to be a plausible interpretation of act-functions such that (a) the interpretation is can be specified independently of how an *ordinary* agent represents her present decision situation, and (b) $\mathcal{B}el$ and $\mathcal{D}es$ are plausible models of her credences and utilities respectively. In the limiting case, we may be able to find an interpretation which satisfies (a) and (b) for *ideally rational* agents, but that won't be of great use in helping us determine $\mathcal{B}el$ and $\mathcal{D}es$ for the ordinary person on the street who is, in the relevant respects, vastly less than ideal.

(I am assuming, here and throughout, that how an agent represents her decision situation can be understood ultimately in terms of her credences. If so, the circularity of characterising credences in terms of $\succcurlyeq$ defined over act-functions is evident. For the purposes of the present dialectic, I do not think that this assumption is unreasonable: it may turn out that the assumption is mistaken, and that representations are not reducible to credence states but some other vaguely doxastic notion—however, I doubt that many proponents of characterisational representationism would be happy to accept a view wherein credences are defined in terms of preferences, which in turn characterised in terms of beliefs/acceptance/opinions/etc.)

Consider again the example with Jill. The upshot of the following discussion is not altered if we characterise Jill's options as *acts* or as *intentions to act*, so to simplify I will usually write just in terms of acts. (Alternatively, we could take an *intention to* ϕ as a kind of mental act.) There is, as a matter of fact, a range of things that Jill can do, behaviours which from a purely physical perspective are within her capacity. For instance, she can *push the button* by extending her arm with fingers pointed in a particular direction, or she can *leave the button alone* by keeping her arms at bay. Each of these has a particular *causal profile*: for each different state that the world might be in, each leads to some outcome or another. From a purely physical-behavioural perspective, then, both *push the button* and *leave the button alone* (and presumably any other act she might engage in) can be associated with unique act-functions defined on some appropriate $\mathcal{S}$ and $\mathcal{O}$, which we will suppose have been given to us for free. (That is, we will ignore the complications mentioned above.)

More specifically, suppose that $\{E_1, E_2, E_3\}$ is a partition of $\mathcal{S}$, and $(E_1, o_1 | E_2, o_2 | E_3, o_3)$ accurately represents the causal profile of, and is therefore associated with, *push the button*; while $(E_1, o_3 | E_2, o_2 | E_3, o_1)$ accurately represents the causal profile of, and is so associated with, *leave the button alone*. Of course, the actual causal profiles of these two behaviours will be much more complicated, but for now, this simplifying fiction does no harm to the example—I will return to this point shortly. We know that Jill did not choose to *push the button*, so applying Savage's conception of preference,

$$(E_1, o_3 | E_2, o_2 | E_3, o_1) \succcurlyeq (E_1, o_1 | E_2, o_2 | E_3, o_3)$$

Assuming $\succcurlyeq$ satisfies Savage's other conditions, then, Jill will be represented as an expected utility ($\mathcal{EU}$) maximiser with credences $\mathcal{B}el$ on $\mathcal{E}$ and utilities $\mathcal{D}es$ on $\mathcal{O}$ such that:

$$\mathcal{E}((E_1, o_3 | E_2, o_2 | E_3, o_1)) \geq \mathcal{EU}((E_1, o_1 | E_2, o_2 | E_3, o_3))$$

Now, we might suppose that Jill is an entirely rational decision-maker: she is extremely mathematically gifted and has perfect introspective access to her credences and utilities, and always chooses the option with the highest expected utility, *given* the way she takes the world to be. The problem, of course, is that if she *misrepresents* the causal profiles of her two options, then the $\mathcal{B}el$ and $\mathcal{D}es$ functions that we arrive at *via* Savage's representation theorem will *ipso facto* be inaccurate. If Jill mistakenly thinks that *push the button* has the causal profile $(E_1, o_3 | E_2, o_2 | E_3, o_1)$, for example, and that *leave the button alone* has the profile $(E_1, o_1 | E_2, o_2 | E_3, o_3)$, then $\mathcal{B}el$ and $\mathcal{D}es$ will not be accurate models of her mental states. From her mistaken perspective, the expected utility of $(E_1, o_1 | E_2, o_2 | E_3, o_3)$ is greater than $(E_1, o_3 | E_2, o_2 | E_3, o_1)$—that is why she chose not to push the button![79]

The problem, of course, is that Jill is being modelled as knowing exactly what acts are available to her, and what the causal profiles of each of those acts are in fact like. But such a model of Jill's decision-making is almost certain to misrepresent whenever there is a mismatch between the actual causal profile of the act (expressed in terms of $\mathcal{S}$ and $\mathcal{O}$) and how she *conceives of* that act's pattern of possible outcomes. Another way in which this kind of misrepresentation might come about would be if Jill did not recognise that,

[79] In his (1973, 249ff, 254), Amartya Sen discusses a related case, wherein observation of an agent's preferences leads to a misrepresentation of her utilities (see also Sen 1993, 501). In Sen's case, a subject represents her options accurately, but is assigned the wrong utilities because it is falsely assumed that she maximises expected utility when in fact she "[follows] a moral code [while] suspending the rational calculus" (251). As he puts it, "People may be induced by social codes of behaviour to act *as if* they have different preferences from what they really have" (258).

say, *push the button* was an option available for her to choose—she may well have chosen *leave the button alone*, had she been aware of the possibility.

Generalising: if the set of act-functions over which $\succcurlyeq$ is defined is supposed to model the range of acts (or intentions to act) *actually* available to S, by virtue of characterising the *actual* causal profiles of those acts (intentions), then $\mathcal{B}el$ and $\mathcal{D}es$ are all but guaranteed to misrepresent S's credences and utilities inasmuch as S misrepresents what options are available to her and/or what outcomes they might lead to. To be sure, a set of act-functions *can* be used to represent a decision situation as it really is, but if the aim is to arrive at descriptively plausible $\mathcal{B}el$ and $\mathcal{D}es$ functions, then they should only be used to represent the situation as the agent represents it.[80]

Such misrepresentation would not arise—at least, assuming that the agent is in fact an expected utility maximiser—if there were a way to *guarantee* the following two conditions:

(1) Each of the act-functions over which $\succcurlyeq$ is defined accurately models both some act's actual causal profile *and* how the agent conceives of its profile

(2) The agent knows (i.e., with certainty) exactly which acts she can and cannot perform, if she were to intend as such

Perhaps, as was noted above, if we understand preferences as relations over *intentions to act*, then we might be on solid ground in supposing that rational agents have full and reliable access to what options are available to her—though (2) still seems implausible for non-ideal agents. But moreover, regardless of whether we take the agent to be deciding between acts or intentions, it's doubtful that how an agent might represent the causal profiles of the things she might do will *always* coincide with the facts of matter.

There may be a special case with ideally rational agents, where we could construct $\mathcal{A}$ in such a manner so as to guarantee both (1) and (2). That is, suppose that states are *dependency hypotheses* (as characterised in §5.2.2), and that ideally rational agents are always fully aware of exactly what acts (or intentions to act) are available for choice.

[80] See (Hausman 2000) for a brief articulation of roughly the same point as applied to revealed preference theory; and (Sen 1993, 502) for a related discussion that distinguishes between 'extensional' and 'intentional' (*sic*) specifications of options. Hausman asserts that "The inverse inference from choice to preference depends … on premises concerning beliefs. Indeed, opposite beliefs and preferences may lead to exactly the same choice" (103). He does not, however, offer a reconciliation of this with the common folklore that representation theorems like Savage's *demonstrate* that some patterns of choices can *only* be the result (*via* expected utility maximisation) of a unique set of credences and utilities. The reconciliation is afforded by recognition of the fact that Savage's representation is only unique *given* a choice of $\mathcal{O}$ and $\mathcal{S}$, and *given* a particular correspondence of act-functions to the actual objects of choice. The mistake is to think that act-functions pick out objects of choice by describing their *actual* patterns of outcomes.

Since every act's outcomes are built into the very specification of each dependency hypothesis, it is plausible to suppose that a function from a set $\mathcal{S}$ of dependency hypotheses to $\mathcal{O}$ could accurately represent not only the act's *actual* causal profile but also at least one way that an ideally rational agent might conceptualise that act's pattern of possible outcomes. We would, of course, still need an appropriate and objective specification of $\mathcal{O}$, and a representation theorem that is well-suited for the use of dependency hypotheses as states (we have seen that Savage's is not), but there are bigger problems here—namely, that this special case will not help us much with non-idealised subjects.

The ideally rational agent never makes logical mistakes, and she is able to keep in mind the full range of dependency hypotheses and consider the relative likelihoods of such things. The ordinary subject cannot do such things. Suppose our decision situation is such that there are only five available acts and four possible outcomes—in which case there are $4^5 = 1024$ dependency hypotheses to consider (and 2^{1024} events). For any *realistic* decision situation, the range of dependency hypotheses is vast and each one is extraordinarily complicated. To expect of the average agent that they could *accurately* represent each act's pattern of possible consequences in this manner is to expect far too much—we don't even come close to representing our decision situations in this way.[81] It is rare enough that decision problems are formulated with states or events which are genuinely act-independent and outcome-functional (which they must be if the states are dependency hypotheses). It is even rarer that *all* the relevant possibilities are taken into account, with each act's actual and complete causal profile being faithfully represented.

The ordinary agent also takes a very coarse-grained conception of pattern of outcomes that her acts (or intentions to act) might have, and the exact nature of that conception seems to be highly variable. Sometimes factors relevant to a decision are simply forgotten about or ignored for whatever reason, leading to variations choices between acts despite no change in the underlying credences and utilities. There are also more systematic phenomena to consider here. For example, it is very plausible on empirical grounds that what aspects of an act's *known outcomes* are salient to a decision-maker is highly context-dependent (Kahneman and Tversky 1979, Tversky and Kahneman 1981, Dietrich and List 2013). A healthcare worker might know that giving a population of 1000 terminally ill patients a particular treatment will cure 75% of them but kill the rest, and in one context focus on the positive aspect of the outcome (750 lives saved) and so proceed with the treatment, while in another context focus on the negative (250 killed) and so choose against it. To account for this variation, we don't need to posit that the agent's underlying utilities for the relatively specific outcome *save 750 and kill 250* changes from context to context. Instead, what seems to happen is that she attaches a high utility to the more

[81] Letting the $\mathcal{S}$ be a collection of dependency hypotheses also only highlights the problems discussed in §5.3. There are many more propositions that we can have credences towards than can be expressed as disjunctions of dependency hypotheses.

coarse-grained prospect of saving 750 people, a low utility to killing 250 people, and contextual factors cause her to represent the act's outcome in one of these two different ways—thus leading to different patterns of preferences regarding the act dependent on context without a change in her credences or utilities.

Savage himself seemed well aware of these issues—hence he distinguished "small world" decision problems from "grand world" problems (see his 1954, 16, 82ff), where the latter is essentially a decision situation modelled such that *all* relevant distinctions between states, outcomes, and acts have been made. The grand world representation of a decision-making context makes use of incredibly fine-grained states and outcomes, while small world representations rely on less specific ways of carving up $\mathcal{S}$ and $\mathcal{O}$. Ordinary subjects, Savage realised, could not hope to contemplate a grand world decision problem, and instead relied on much more coarse-grained (or small world) conceptions of their situation. In Savage's system, act-functions might represent acts within a small world or a grand world conceptualisation. But the distinction between small and grand worlds will not help us with our present problems, as it seems unlikely that we could know *how* coarse-grained a conception the agent has taken of her circumstances without having prior access to her mental states. The only *objective* representation of a decision situation is the maximally fine-grained one, which ordinary agents are incapable of conceptualising in all its detail.

From an outsider's perspective, it may be possible to specify what acts are within *S*'s physical capabilities, and with a detailed enough knowledge of physics, exactly what the causal profile of each such act actually is—all without peaking inside *S*'s head. But this is not the kind of information we require if we are going to model why *S* made the choice she did. To understand her choices, it will not do to model her range of options as they *actually* are, *if* how they actually are is distinct from how she *takes them to be*. And this requires access to how *S* represents her present decision situation, which isn't the kind of information we can have from the outside.

5.5 Summary

I have argued that the biggest concerns for theorems within the Savage paradigm originate with the use of act-functions as the basic relata of $\succcurlyeq$—ironically so, given that it is because $\succcurlyeq$ is defined over act-functions that Savage's framework is so frequently used. Many in the behavioural sciences find this feature of the framework particularly attractive. Indeed, Fishburn criticises Jeffrey's representation theorem—where $\succcurlyeq$ is defined over an algebra of propositions—on the basis that it "blurs the often useful distinctions among acts, consequences, and other entities that appear in other [multi-set] theories" (Fishburn 1981, 186, cf. Bolker 1967, 335).

Part of this attraction is due to a latent methodological behaviourism which still persists today, according to which it will not do to simply *ask* a subject what her credences and utilities are: the only *accurate* measure of such things can come from observation of her choices between acts. (See Gul and Pesendorfer 2008 for a recent defence of this idea.) But even foregoing methodological behaviourism, there is the widely-held idea that decision theory is about *acts*, and so $\succcurlyeq$ must be construed behaviourally and its basic relata formalised accordingly. However, we have seen that any representation theorem which appeals to act-functions will be sub-optimal for the purposes of characterisational representationism. To close this chapter, I will summarise the issues raised in this chapter by their relation to the desiderata established in §3.4.5.

There are, first of all, a number of issues which centre on the apparently crucial appeal to imaginary act-functions. On the one hand, it seems unlikely that any interesting representation result can be achieved without imposing some rather strong structural requirements on the space of act-functions, $\mathcal{A}$—requirements which seem incompatible with taking $\mathcal{A}$ to be a formal representation of a space of available *acts*, $\mathcal{A}'$. This makes the behavioural interpretation of $\succcurlyeq$ highly problematic. Formally, $\succcurlyeq$ must be defined on a set with structural properties simply not possessed by $\mathcal{A}'$, the supposed basic objects of our behavioural preferences. This is in conflict with desideratum (**1a**), that a theorem's preference conditions should be *satisfiable tout court*. On the other hand, it may be possible to re-interpret $\mathcal{A}$ as representing one of the following:

(i) A set of *imaginable* acts
(ii) A set of (im)possible patterns of outcomes
(iii) The union of $\mathcal{A}'$ with a set $\mathcal{A}^*$ of purely fictional entities

However, each option gives rise to issues relating to desideratum (**1**) more generally, and (i) and (ii) additionally require giving up on any non-intentional interpretation of $\succcurlyeq$ (which leads to issues with desideratum (**5**)). It may turn out, however, that the notion of *coherent extendibility* can be of service to characterisational representationism here—though it is not obvious if what results will be in conflict with (**3**).

More worrying is the fact that Savagean theorems seem incapable of satisfying desideratum (**2**), that $\mathcal{B}el$ and $\mathcal{D}es$ should provide plausible models of the relevant agents' total credence and utility states. Savage's own theorem is already problematic on this front by virtue of being a CEU theorem (see desideratum (**2d**)). However, there are still deeper worries here. The $\mathcal{B}el$ and $\mathcal{D}es$ functions that *any* Savage-like theorem might supply us with have impoverished domains, putting them in conflict with desideratum (**2c**). Furthermore, the domains of $\mathcal{B}el$ and $\mathcal{D}es$ are wholly disjoint, giving rise to a conflict with (**2a**).

Furthermore, although we did not discuss the point in any detail, a glance at any of the examples given in Appendix B will highlight that contemporary Savage-like theorems also tend to come into conflict with (**2b**), and (**2d**), at least where $\mathcal{S}$ is a partition of some possibility space. Most Savage-like theorems require that $\mathcal{B}el$ has at least as much structure as a Choquet capacity, if not a probability function, and such functions are not well-suited for the representation of ordinary agents' credences. One might try to alleviate these worries by letting $\mathcal{S}$ include impossible states (as noted in §4.3), but doing so would only put more pressure on the already problematic interpretation of Savage's act-functions—for instance, the issues surrounding imaginary act-functions will only be magnified where $\mathcal{S}$ includes both possible and impossible states of affairs.

Finally, it seems that the interpretation of act-functions must be given partially in terms of how agents represent the circumstances they find themselves in, raising worries about whether desideratum **(4)** can be satisfied by any theorem which appeals to act-functions. That is, it seems that the basic notions involved in the interpretation of any such theorem cannot be understood independently of the agent's overall doxastic state, nor of what she cares about. A failure to satisfy **(4)** also implies the failure to satisfy the naturalistic desideratum **(5)**.

It seems, therefore, safe to say that Savage's framework stands at odds with characterisational representationism, and even more so with the naturalisation project—despite its origins in mid-twentieth century behaviourism. Another kind of representation theorem will need to be found.

Chapter six

Lottery-Based and Monoset Theorems

This chapter focuses on two rather different kinds of representation theorem, each of which raise distinctive issues in relation to characterisational representationism. We begin with what I will call *lottery-based* theorems (§6.1), and follow with the *monoset* theorems of Ethan Bolker and Richard Jeffrey (§6.2).

6.1 Lottery-based theorems

The class of lottery-based theorems comprises those which appeal to what have become known as *extraneous scaling probabilities*—in a sense to be made more precise below, these theorems require us to essentially plug in some credence values by hand, rather than deriving them from preferences. One of the two main complaints that I will draw in this section results from this fact, and has often been raised over the past several decades.

The earliest lottery-based theorem originates with John von Neumann and Oskar Morgenstern's (henceforth: VNM) seminal *Theory of Games and Economic Behaviour* (1947), which formed the basis of the theorem developed by Anscombe and Aumann (henceforth: AA) in their 'A Definition of Subjective Probability' (1963). What is now known as the *AA framework* is the basis for a large number of recent representation theorems. For example, the AA framework is used in the theorems of (Fishburn 1970, 1973, 1975, 1982), (Hazen 1987), (Gilboa and Schmeidler 1989), (Blume, Brandenburger *et al.* 1991), (Maccheroni, Marinacci *et al.* 2006), (Seo 2009), (Neilson 2010), and (Schneider and Nunez 2015). The theorems of (Pratt, Raiffa *et al.* 1965), (Balch and Fishburn 1974), (Armendt 1986), and a number of others are based around their own distinctive systems, but like VNM's and AA's theorems, each appeals to extraneous scaling probabilities.

6.1.1 Anscombe and Aumann's theorem

Before I outline the VNM and AA theorems, it will be helpful to begin with an informal characterisation of *lotteries*, which form the intended interpretation of the basic relata of both theorems' $\succcurlyeq$. As AA describe it, a lottery is:

> … a device for deciding which prize in [a set of outcomes $\mathcal{O}$] you will receive, on the basis of a single observation that records which one of a set of mutually exclusive and exhaustive uncertain events took place. (1963, 200)

AA imagine their agent as having a choice between a number of free *lottery tickets*, where the possession of any such ticket enters her into a draw for one of a finite number of prizes, the draw itself being dependent on some way that the world may turn out to be. It is, of course, implicitly assumed that the agent is under no illusions or misconceptions regarding the prize conditions for any of the lotteries she may choose to enter. (This is analogous to the assumptions required of agents and acts noted in §5.4.)

AA distinguish two kinds of lotteries on the basis of the kind of uncertainty involved in the lottery's conditions. The first kind they refer to as *roulette lotteries*; these are lotteries where the *objective chances* associated with each of the prizes being won are known (with certainty) to the agent. So, for instance, assume that a 38-pocket roulette wheel is spun, and let $\{P_1, \ldots, P_{38}\}$ be *the ball lands in pocket 1*, …, *the ball lands in pocket 38*. The ball must land in one pocket or another, so P_1 through to P_{38} are mutually exclusive and exhaustive. AA then imagine a lottery with prize o_1 if P_1 turns out to be true, o_2 if P_2 turns out to be true, and so on. They assume that any ordinary decision-maker will know that each proposition in $\{P_1, \ldots, P_{38}\}$ has an objective chance 1/38 of coming true, and will set her credences accordingly (cf. the 'Principal Principle', Lewis 1980b). Other kinds of roulette lotteries might include those based on the toss of a fair coin, the roll of an n-sided die, or the occurrence of a quantum event with a known probability distribution.

The second kind of lottery is a *horse lottery*, wherein the objective chances associated with each of the prizes being won aren't known—either due to an ignorance of the chances on the decision-maker's behalf, or because there are *no* objective chances associated with the lottery's win conditions.

> On the other hand, [unlike roulette lotteries,] it is possible that chances [for a lottery's outcomes] cannot be associated with the uncertain events in question, or that the values of such chances are unknown; for example, this would be so if we were observing a horse race. (1963, 200)

AA's idea is that while ordinary, rational subjects can be presumed to know the objective chances associated with roulette wheels, they don't know the chances associated with each horse in a race coming first. Given a five-horse race, {*Horse one wins*, …, *Horse five wins*} partitions the relevant space of possibilities, but most agents would not think that each proposition has a 1/5 chance of becoming true: some of the horses are, presumably, objectively better runners than their competitors. Other kinds of horse lotteries would include those based on, say, whether it rains in Sydney on the 15th of May, 2018; or whether a jar contains n jelly beans.

The VNM theorem is based solely around preferences over roulette lotteries. AA regard this situation as unsatisfactory. Because roulette lotteries take the associated credence values as extraneously *given*, there is no sense in which they can help us to characterise *what it is* to have such-and-such credence states. The goal of AA's 'definition of subjective probability' is to use preferences over both kinds of lotteries to derive credence values for the propositions used in the formulation of horse lotteries—that is, propositions such that their objective chances are *not* known. But before we can understand AA's proposal, we will first need to look at VNM's theorem.

Two definitions are needed to begin with. The first gives us the formalisation of a roulette lottery:

Definition 6.1: Lottery-function

A function $\mathcal{L}\colon \mathcal{O} \mapsto [0, 1]$ is a *lottery-function* iff $\mathcal{L}(o) = 0$ for all but a finite number of $o \in \mathcal{O}$, and $\sum_{\mathcal{O}} \mathcal{L}(o) = 1$

A lottery-function on $\mathcal{O}$ is supposed to represent a roulette lottery which associates every outcome in $\mathcal{O}$ with a particular chance, such that the chances sum to one. There is no mention of propositions to be found in Definition 6.1, but they figure essentially in the interpretation of any given lottery-function. The idea is that since the objective chances are assumed to be known, outcomes can be associated with chance values *directly* rather than being associated with the propositions which have those values. We must understand chance values as attaching to some member of a set of mutually exclusive and exhaustive propositions, however—without this, a lottery-function would just be a meaningless pairing of outcomes with numbers. We can also now precisely define a *lottery-based theorem* as any theorem in which lottery-functions can be found amongst the *basic* formal elements of the theorem.

Next, to formalise the set of all possible roulette lotteries with outcomes taken from a set of outcomes $\mathcal{O}$, we will need the notion of a mixture set:

Definition 6.2: Mixture set

A set $\mathcal{M}$ is a *mixture set* iff, for any $x, y \in \mathcal{M}$ and any $\lambda \in [0, 1]$, we can associate another element of $\mathcal{M}$, to be designated (x, λ, y), such that, for all $x, y \in \mathcal{M}$ and all $\lambda, \mu \in [0, 1]$,

(i) $(x, 1, y) = x$

(ii) $(x, \lambda, y) = (y, 1 - \lambda, x)$

(iii) $((x, \lambda, y), \mu, y) = (x, \lambda\mu, y)$

We can now designate the set of all lottery-functions on $\mathcal{O}$ as the mixture set $\mathcal{M}_{\mathcal{O}}$, under the interpretation that $(\mathcal{L}_1, \lambda, \mathcal{L}_2)$ is the lottery-function $\mathcal{L}_3$ such that:

$$\mathcal{L}_3(o) = \lambda.\mathcal{L}_1(o) + (1 - \lambda).\mathcal{L}_2(o), \text{ for all } o \in \mathcal{O}$$

Suppose that $\succcurlyeq$ is defined on $\mathcal{M}_\mathcal{O}$, where $\succcurlyeq$ can be understood either behaviourally as a disposition to choose one lottery ticket over another, or mentalistically as a preference to be holding one ticket over another. The VNM theorem is then quite straightforward, with three simple, necessary preference conditions—namely, for all $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3 \in \mathcal{M}_\mathcal{O}$,

VNM1 $\succcurlyeq$ on $\mathcal{M}_\mathcal{O}$ is a weak ordering

VNM2 If $\mathcal{L}_1 \succ \mathcal{L}_2$ and $0 < \lambda < 1$, then $(\mathcal{L}_1, \lambda, \mathcal{L}_3) \succ (\mathcal{L}_2, \lambda, \mathcal{L}_3)$

VNM3 If $\mathcal{L}_1 \succ \mathcal{L}_2 \succ \mathcal{L}_3$, then for some $\lambda, \gamma \in (0, 1)$, $(\mathcal{L}_1, \lambda, \mathcal{L}_3) \succ \mathcal{L}_2$ and $\mathcal{L}_2 \succ (\mathcal{L}_1, \gamma, \mathcal{L}_3)$

VNM are then able to prove the following:

Theorem 6.1: von Neumann and Morgenstern's theorem

$<\mathcal{M}_\mathcal{O}, \succcurlyeq>$ satisfies **VNM1–VNM3** iff there exists a real-valued function $\mathcal{D}es$ on $\mathcal{M}_\mathcal{O}$ such that, for all $\mathcal{L}_1$, $\mathcal{L}_2$, $(\mathcal{L}_1, \lambda, \mathcal{L}_2) \in \mathcal{M}_\mathcal{O}$,

(i) $\mathcal{L}_1 \succcurlyeq \mathcal{L}_2$ iff $\mathcal{D}es(\mathcal{L}_1) \geq \mathcal{D}es(\mathcal{L}_2)$

(ii) For all $\lambda \in [0, 1]$, $\mathcal{D}es((\mathcal{L}_1, \lambda, \mathcal{L}_2)) = \lambda.\mathcal{D}es(\mathcal{L}_1) + (1 - \lambda).\mathcal{D}es(\mathcal{L}_2)$

Furthermore, $\mathcal{D}es$ is unique up to positive linear transformation

Importantly, $\mathcal{D}es$ can be defined on $\mathcal{O}$ by letting $\mathcal{D}es(o) = \mathcal{D}es(\mathcal{L})$, where $\mathcal{L}$ is the *trivial* lottery-function which assigns a chance of 1 to o—the idea being that the lottery $\mathcal{L}$ represents presumably has exactly the same utility as o: it is effectively just a guarantee that o. So, for every lottery $\mathcal{L} \in \mathcal{M}_\mathcal{O}$,

$$\mathcal{D}es(\mathcal{L}) = \textstyle\sum_\mathcal{O} \mathcal{L}(o).\mathcal{D}es(o)$$

We can also extend $\succcurlyeq$ to $\mathcal{O}$ by assuming that for all o, $o^* \in \mathcal{O}$,

$$o \succcurlyeq o^* \text{ iff } \mathcal{D}es(o) \geq \mathcal{D}es(o^*)$$

So much for von Neumann and Morgenstern's theorem, let us now look at Anscombe and Aumann's development.

The basic trick to AA's theorem is a dual application of Theorem 6.1 to *two* preference relations $\succcurlyeq$ and $\succcurlyeq^*$ (defined on disjoint sets of lottery-functions), with the addition of two further preference conditions to connect $\succcurlyeq$ and $\succcurlyeq^*$ together. $\succcurlyeq$ is identical to VNM's relation, and defined on the set $\mathcal{M}_{\mathcal{O}}$. On the other hand, $\succcurlyeq^*$ is supposed to represent preferences between *roulette lotteries with horse lotteries as prizes*, where those horse lotteries have yet more roulette lotteries as prizes.

Formally, horse lotteries are functions from a set $\mathcal{S}$ of states—where these states are understood basically as Savage understood them—into $\mathcal{M}_{\mathcal{O}}$. Let us refer to such functions as *horse-functions*, and let $\mathcal{H} = \{\hbar_1, \hbar_2, \hbar_3, \ldots\}$ be the set of all horse-functions (generally: $\mathcal{H} \subseteq \mathcal{M}_{\mathcal{O}}{}^{\mathcal{S}}$). So, every $\hbar \in \mathcal{H}$ assigns some lottery-function $\mathcal{L} \in \mathcal{M}_{\mathcal{O}}$ to each $s \in \mathcal{S}$. As with Savage's $\mathcal{A}$, $\mathcal{H}$ will include *constant functions* which assign the same $\mathcal{L}$ to every state in $\mathcal{S}$. If $\hbar$ assigns a trivial lottery-function to each state, then it is *in effect* just one of Savage's act-functions. However, the 'prizes' associated with a horse-function may include *any* lottery-function in $\mathcal{M}_{\mathcal{O}}$.

Let $\mathcal{M}_{\mathcal{H}}$ be the set of all lottery-functions defined on $\mathcal{H}$. AA's preference conditions are much simpler than Savage's, though this is offset by the far greater complexity involved in the characterisation of $\mathcal{M}_{\mathcal{O}}$ and $\mathcal{M}_{\mathcal{H}}$. The first two preference conditions are easy to state:

AA1 $\succcurlyeq$ on $\mathcal{M}_{\mathcal{O}}$ and $\succcurlyeq^*$ on $\mathcal{M}_{\mathcal{H}}$ satisfy **VNM1–VNM3**

AA2 There are $o, o^* \in \mathcal{O}$ such that $o \succ o^*$ and for all $o^+ \in \mathcal{O}$, $o \succcurlyeq o^+ \succcurlyeq o^*$

Given Theorem 6.1, we know from **AA1** that there will be *two* utility functions $\mathcal{D}es$ and $\mathcal{D}es^*$ which T-represent $\succcurlyeq$ on $\mathcal{M}_{\mathcal{O}}$ and $\succcurlyeq^*$ on $\mathcal{M}_{\mathcal{H}}$ respectively. **AA2** is included for mathematical ease; it simply asserts that there are at least two distinct outcomes o and o' in $\mathcal{O}$ such that o is the most desired outcome, o' is the least desired outcome, and the agent is not indifferent between them.

The next two preference conditions will require some further notation to express. First, we will represent an arbitrary lottery-function $\mathcal{L}$, with prizes $x_1, \ldots, x_m$ in either $\mathcal{O}$ or $\mathcal{H}$, as $\langle \lambda_1, x_1 | \ldots | \lambda_m, x_m \rangle$, where $\lambda_1, \ldots, \lambda_m \in (0, 1]$ and $\sum \lambda_i = 1$. Thus, $\langle \lambda_1, x_1 | \ldots | \lambda_m, x_m \rangle$ is supposed to represent the roulette lottery which has a λ_1 chance of resulting in x_1, a λ_2 chance of resulting in x_2, and so on. The trivial lottery-function which assigns a chance of 1 to x being won is then represented $\langle 1, x \rangle$. Secondly, we will represent an arbitrary horse-function $\hbar$ as follows: if $\hbar$ may result in exactly n different lottery-functions $\mathcal{L}_1, \ldots, \mathcal{L}_{\mathrm{n}}$ at the events $E_1, \ldots, E_n$ respectively, then we will represent it $[\![\mathcal{L}_1, \ldots, \mathcal{L}_{\mathrm{n}}]\!]$. Thus, $[\![\mathcal{L}_1, \ldots, \mathcal{L}_{\mathrm{n}}]\!]$ is supposed to designate the horse lottery with the prize $\mathcal{L}_1$ if any $s \in E_1$ is true, $\mathcal{L}_2$ if any $s \in E_2$ is true, and so on, where $\{E_1, \ldots, E_n\}$ is a partition of $\mathcal{S}$.[82]

[82] Because of their similarity to Savage's act-functions, a horse-function can also be represented along the same lines as act-functions were in Chapter 5; that is, with $\hbar$ being denoted $(E_1, \mathcal{L}_1 | \ldots | E_{\mathrm{n}}, \mathcal{L}_{\mathrm{n}})$ iff $\hbar(s) = \mathcal{L}_1$ for all $s \in E_1$, $\hbar(s) = \mathcal{L}_2$ for all $s \in E_2$, and so on. I have altered the notational scheme to improve the legibility **AA4**, which would be rendered (even more) opaque under the earlier scheme.

We can now state AA's final two preference conditions, which are supposed to hold for all lottery-functions in $\mathcal{M}_{\mathcal{H}}$:

AA3 If $\mathcal{L}_{\mathrm{i}} \succcurlyeq \mathcal{L}_{\mathrm{i}}'$, then $\langle 1, [\![\mathcal{L}_1, \ldots, \mathcal{L}_{\mathrm{i}}, \ldots, \mathcal{L}_n]\!]\rangle \succcurlyeq^* \langle 1, [\![\mathcal{L}_1, \ldots, \mathcal{L}_{\mathrm{i}}', \ldots, \mathcal{L}_n]\!]\rangle$

AA4 $\langle \lambda_1, [\![\mathcal{L}^1{}_1, \ldots, \mathcal{L}^1{}_n]\!] \mid \ldots \mid \lambda_m, [\![\mathcal{L}^m{}_1, \ldots, \mathcal{L}^m{}_n]\!]\rangle \sim^* \langle 1, [\![\langle \lambda_1, \mathcal{L}^1{}_1 \mid \ldots \mid \lambda_m, \mathcal{L}^m{}_1\rangle, \ldots, \langle \lambda_1, \mathcal{L}^n{}_1 \mid \ldots \mid \lambda_m, \mathcal{L}^m{}_n\rangle]\!]\rangle$

AA3 says that if two horse-functions $\hbar$ and $\hbar'$ are identical except that at some event E, and $\hbar(s) = \mathcal{L}_{\mathrm{i}}$ and $\hbar'(s) = \mathcal{L}_{\mathrm{i}}'$ for all $s \in E$, then the agent′s preferences between $\hbar$ and $\hbar'$ are determined by her preferences between $\mathcal{L}_{\mathrm{i}}$ and $\mathcal{L}_{\mathrm{i}}'$. While all but opaque to merely human eyes, **AA4** is picturesquely described by AA as saying that "if the prize you receive is to be determined by both a horse race and a spin of a roulette wheel, then it is immaterial whether the wheel is spun before or after the race" (1963, 201).

AA note that their **AA3** implies that if o and o' are the most and least preferred outcomes respectively, then the trivial lottery-function in $\mathcal{M}_{\mathcal{H}}$ which is guaranteed to result in o (i.e., $\langle 1, [\![\langle 1, o\rangle]\!]\rangle$) will be the most preferred element of $\mathcal{M}_{\mathcal{H}}$, while the trivial lottery-function $\langle 1, [\![\langle 1, o'\rangle]\!]\rangle$ guaranteed to result in o' will be the least preferred element of $\mathcal{M}_{\mathcal{H}}$. They therefore propose to normalise $\mathcal{D}es^*$ on $\mathcal{M}_{\mathcal{H}}$ by letting $\mathcal{D}es^*(\langle 1, [\![\langle 1, o\rangle]\!]\rangle) = 1$, and $\mathcal{D}es^*(\langle 1, [\![\langle 1, o'\rangle]\!]\rangle) = 0$. Likewise, they set $\mathcal{D}es(\langle 1, o\rangle) = 1$ and $\mathcal{D}es(\langle 1, o'\rangle) = 0$. Given this, we can now state AA′s theorem:

Theorem 6.2: Anscombe and Aumann

If $\succcurlyeq$ on $\mathcal{M}_{\mathcal{O}}$ and $\succcurlyeq^*$ on $\mathcal{M}_{\mathcal{H}}$ satisfy **AA1–AA4**, then there exists two normalised utility functions, $\mathcal{D}es\colon \mathcal{M}_{\mathcal{O}} \mapsto \mathbb{R}$ and $\mathcal{D}es^*\colon \mathcal{M}_{\mathcal{H}} \mapsto \mathbb{R}$, and a unique probability function $\mathcal{B}el\colon \boldsymbol{\mathcal{E}} \mapsto [0, 1]$, such that for all $\mathcal{L}_1, \mathcal{L}_2 \in \mathcal{M}_{\mathcal{O}}$ and all $\mathcal{L}'_1, \mathcal{L}'_2, [\![\mathcal{L}_1, \ldots, \mathcal{L}_{\mathrm{n}}]\!] \in \mathcal{M}_{\mathcal{H}}$,

(i) $\mathcal{L}_1 \succcurlyeq \mathcal{L}_2$ iff $\mathcal{D}es(\mathcal{L}_1) \geq \mathcal{D}es(\mathcal{L}_2)$

(ii) $\mathcal{L}'_1 \succcurlyeq^* \mathcal{L}'_2$ iff $\mathcal{D}es^*(\mathcal{L}'_1) \geq \mathcal{D}es^*(\mathcal{L}'_2)$

(iii) $\mathcal{D}es^*([\![\mathcal{L}_1, \ldots, \mathcal{L}_{\mathrm{n}}]\!]) = \sum_i^n \mathcal{B}el(E_{\mathrm{i}}).\mathcal{D}es(\mathcal{L}_{\mathrm{i}})$

Because $\succcurlyeq$ and $\succcurlyeq^*$ are defined on disjoint sets, AA point out that there would be no harm done in restating their theorem using a single preference relation $\succcurlyeq^+$ defined on $\mathcal{M}_{\mathcal{O}} \cup \mathcal{M}_{\mathcal{H}}$, by simply letting $\succcurlyeq^+$ equal $\succcurlyeq$ on $\mathcal{M}_{\mathcal{O}}$ and $\succcurlyeq^*$ on $\mathcal{M}_{\mathcal{H}}$.

6.1.2 Critical discussion

The similarity between lottery-functions and horse-functions on the one hand, and act-functions on the other, should make it clear that many of the same issues which arose for Savage's theorem have close analogues for AA's and similar theorems. I will briefly outline these, before moving on to the problems that arise specifically from appealing to lottery-functions. Before we begin, though, it's worth noting that the interpretation of

VNM's or AA's basic objects of preference as *lotteries* is not forced upon us by their formalism. Indeed, it is fairly common to treat each of their 'lotteries' as a variation on the basic idea of an act-function—after all, both AA's 'lotteries' and Savage's act-functions are ultimately just pairings of uncertain events with outcomes. The exact interpretation we assign to $\mathcal{M}_\mathcal{O}$, $\mathcal{M}_\mathcal{H}$, and $\mathcal{H}$ is largely immaterial for the purposes of my critical discussion, so I will follow AA in describing the relata of $\succcurlyeq$ as lotteries.

First of all, AA make essential use of trivial lottery-functions and constant horse-functions, and this gives rise to something very much like the problem of constant acts as it appears in Savage's system (§5.2)—a lottery-function $\langle 1, [\![\langle 1, o\rangle]\!]\rangle$ *is*, for all intents and purposes, just a constant act-function $\underline{o}$. Furthermore, an analogue of the problem discussed in §5.4 can be raised for the interpretation of each of $\mathcal{M}_\mathcal{O}$, $\mathcal{M}_\mathcal{H}$, and $\mathcal{H}$: to whatever extent these sets represent lotteries (or any other objects of choice), they must represent them *as the agent takes them to be* rather than just *as they in fact are*, else the derived $\mathcal{B}el$ and $\mathcal{D}es$ functions are all but guaranteed to misrepresent decision-makers' credences and utilities. Finally, there is the question of whether AA's $\mathcal{B}el$ and $\mathcal{D}es$ functions are defined on a domain with the right kind of structure to adequately represent ordinary agents' credences and utilities (§5.3). Given a lottery-based interpretation, the states in $\mathcal{S}$ are no longer required to be act-independent—but they should instead be, in some relevant sense, *lottery-independent*, as well as outcome-independent. This in turn implies that the states in $\mathcal{S}$ are independent with respect to each of the propositions which characterise the win conditions for any of the lotteries in $\mathcal{M}_\mathcal{O}$. Thus, $\mathcal{B}el$ can only model of credences *specifically with respect to* those propositions towards which the objective credences aren't known.[83]

A complete model of the agent's credences would presumably take the form of an extension of $\mathcal{B}el$—call it $\mathcal{B}el^+$—which equals $\mathcal{B}el$ on $\mathcal{E}$ but also represents credence values for those propositions not in $\mathcal{E}$ for which the objective chances are known. Call the set of propositions towards which the agent knows the objective chances $\mathcal{P}_\mathcal{O}$; because $(\mathcal{E} \cap \mathcal{P}_\mathcal{O}) = \emptyset$, the imagined extension of $\mathcal{B}el$ should not be problematic. However, we will need some *independent* means of fixing $\mathcal{B}el^+$ on $\mathcal{P}_\mathcal{O}$—and this is something which cannot be given by any lottery-based theorem.

We are led, then, to the most commonly recognised problem with lottery-based theorems in relation to characterisational representationism: the credences values associated with a certain large subset of propositions ($\mathcal{P}_\mathcal{O}$) must be given independently. AA implicitly assume that the objective chances associated with certain propositions (such as *the toss of a fair coin will land heads* and *the ball on the roulette wheel will land in the first*

[83] It should also be noted that while most lottery-based theorems *do* make use of AA's horse-functions (i.e., functions from $\mathcal{S}$ to $\mathcal{M}_\mathcal{O}$), this is not essential to the lottery-based framework—for example, Armendt's (1986) theorem is lottery-based, but has preferences which are defined between arbitrary propositions and lottery-functions with propositions as 'prizes'.

pocket) are common knowledge. But an appeal to a lottery-based theorem leaves us without an explanation of where *these* credences come from. At best, a lottery-based theorem could be used to construct a *partial* model of credences—but *only* with respect to events in $\mathcal{E}$ (which is already highly impoverished), and *only if* certain conditions hold of the agent's credences towards the propositions in $\mathcal{P}_{\mathcal{O}}$.

It seems unlikely, however, that any theorem like AA's would even be *useful* for the limited task of characterising credences for the propositions *not* in $\mathcal{P}_{\mathcal{O}}$. In particular, it seems *prima facie* plausible that *if* we were able to given an account of *what it is* to have a credence of x in P, where x might take any value in [0, 1] and P might be any proposition in a rather large set $\mathcal{P}_{\mathcal{O}}$, then that *same* account should apply to any propositions whatsoever. There seems to be no good reason to think that the metaphysics of credences should be *disjunctive*, in the sense of there being one account of credences towards the propositions in $\mathcal{P}_{\mathcal{O}}$, and another for propositions outside of $\mathcal{P}_{\mathcal{O}}$ (cf. §5.3.1).

Compounding this problem is, I think, an even bigger concern. The lottery-functions in $\mathcal{M}_{\mathcal{H}}$ are *incredibly* unintuitive constructions, being supposed to represent roulette lotteries with horse lotteries as prizes, which in turn have roulette lotteries with prizes in $\mathcal{O}$ as prizes! It is *immensely* implausible that anyone's credences and utilities should be characterised primarily in terms of their preferences over such things, if we even *have* preferences over such things. Such an *odd* domain for $\succcurlyeq$ is surely the wrong place to look if we are seeking a plausible basis for characterising credences and utilities. Perhaps a 'disjunctivist' account of credences could be defended, but not if one of those disjuncts involves preferences over $\mathcal{M}_{\mathcal{H}}$. Preferences over *lotteries upon lotteries upon lotteries* just don't seem like the kinds of things that we should want to base an account of credences and utilities upon.

AA themselves offer an inadequate justification for referring to objective chances and preferences over roulette lotteries. In particular, they suggest that the notion of credence might be "even obscurer than chance and that progress [with respect to characterising credences] should preferably be from the more familiar to the less familiar, rather than the other way around" (1963, 203). Thus they propose to define a person's credences in terms of objective chances. It is obvious, though, that AA do *not* reduce the notion of credence to the notion of chance, despite their claims to have done so (e.g., 199-200). It is immensely implausible that an agent's preferences over roulette lotteries are *directly* determined by the objective chances which are associated with those lotteries' win conditions. Rather, just as is the case with horse lotteries, a rational agent prefers one roulette lottery $\mathcal{L}$ over another $\mathcal{L}'$ if her credences are such that $\mathcal{L}$ is subjectively more likely to result in the better outcome. The *only* difference between roulette lotteries and horse lotteries is that in the former case, the objective chances associated with the lottery's win conditions are assumed to be *known* (and the agent's credences are set accordingly). Note, of course, that this assumption is unlikely to be true *in general*: ordinary agents make mistakes, and may fail to accurately represent the chances associated with the lotteries on offer to them.

AA don't trade one obscure notion (*credence*) for a less obscure notion (*chance*)—and if they are right in thinking that the notion of *credence* is obscure, then their theorem does little to remove that obscurity. Characterisational representationism will need a much stronger foundation than a lottery-based theorem. Unfortunately, that means doing without a very great many of the theorems which have been developed over the past few decades.

6.2 Monoset theorems

Each of the theorems considered thus far have been *multiset* theorems—that is, the objects of credence, utility, and preference are formally represented by distinct sets (e.g., in Savage: $\mathcal{E}$, $\mathcal{O}$, and $\mathcal{A}$; and in Anscombe and Aumann: $\mathcal{E}$, $\mathcal{O}$, and $\mathcal{M}_{\mathcal{O}} \cup \mathcal{M}_{\mathcal{H}}$). By contrast, the *monoset* theorem to be considered here has its objects of credence, utility, and preference all drawn from a single set of propositions, $\mathcal{P}$.

The mathematical basis for the theorem that we will now consider was first developed by Bolker (1966, 1967), and its application in decision theory was extensively discussed in Jeffrey (e.g., 1978, 1990).

6.2.1 The Jeffrey-Bolker theorem

We begin with a preference relation $\succcurlyeq$ defined on a σ-algebra of propositions $\mathcal{P}$, where propositions are understood as sets of worlds taken from some infinite space of worlds $\mathcal{W}$. The propositions in $\mathcal{P}$ will be assigned credences and utilities in the final representation. Since $\succcurlyeq$ is not defined on objects of choice (*à la* Savage) but on arbitrary propositions, it is best understood in the mentalistic sense. (For more on this, see also the discussion in Chapter 9.)

As with each of the other theorems we have looked at, monoset theorems involve a number of background structural and non-triviality conditions:

MON0 For all $P \in \mathcal{P}$, if $P \notin \mathcal{N}$, then there exists two non-empty propositions $Q, Q' \in \mathcal{P}$ such that $Q, Q' \notin \mathcal{N}$, $(Q \cap Q') = \emptyset$, and $P = (Q \cup Q')$

MON1 For some $R \in \mathcal{P}$, $R \succ (R \cup \neg R) \succ \neg R$

MON2 $\succcurlyeq$ on $\mathcal{P}$ is a weak ordering

The purely structural condition **MON0** specifies that the set of propositions $\mathcal{P} - \mathcal{N}$ is bottomless, and thus infinite; the set $\mathcal{N}$ to which it refers will be defined below. **MON1** is required if any interesting representation of the agent's preferences is to exist; the proposition R that it mentions will be used to scale $\mathcal{D}es$.

Next, for all relevant propositions,

MON3 If $(P \cap Q) = \emptyset$ and $P \succcurlyeq Q$, then $P \succcurlyeq (P \cup Q) \succcurlyeq Q$

MON3 says something similar to **MON1**, though generalised to all disjoint pairs of propositions. Essentially, it requires that the utility of a disjunction of two incompatible propositions P and Q should sit somewhere weakly between the utilities of P and Q.

The next preference condition is crucially important for the existence of the desired representation; in particular, failure to satisfy this condition results in a probabilistically incoherent $\mathcal{B}el$ function.

MON4 If $P \succcurlyeq P'$ and if $(P \cap Q) = (P \cap Q') = (P' \cap Q) = (P' \cap Q') = \emptyset$, then either (i) $\neg(Q \succ (P' \cup Q) \succcurlyeq (P \cup Q) \succ P \succcurlyeq P' \succ (P' \cup Q') \succcurlyeq (P \cup Q') \succ Q')$, or (ii) $Q \succ (P' \cup Q) \sim (P \cup Q) \succ P \sim P' \succ (P' \cup Q') \sim (P \cup Q') \succ Q'$

The basic role of **MON4** is to similar to that of Savage's condition **SAV5**; namely, it is used to connect $\succcurlyeq$ with a relative credence relation $\succcurlyeq^{b}$ on $\mathcal{P}$ using a variation on Savage's Coherence principle:

Definition 6.3: Monoset coherence
If $P, P', Q \in \mathcal{P}$, and $(P \cap Q) = (P' \cap Q) = \emptyset$, then:

(i) $P \succcurlyeq^{b} P'$ if $(Q \succ (P' \cup Q) \succcurlyeq (P \cup Q) \succ P \succcurlyeq P')$ or $(P' \succcurlyeq P \succ (P' \cup Q) \succcurlyeq (P \cup Q) \succ Q)$

(ii) $P \sim^{b} P'$ if $(Q \succ (P' \cup Q) \sim (P \cup Q) \succ P \sim P')$ or $(P' \sim P \succ (P' \cup Q) \sim (P \cup Q) \succ Q)$

(iii) $P \succ^{b} P'$ if $(Q \succ (P' \cup Q) \succcurlyeq (P \cup Q) \succ P \succcurlyeq P')$ or $(P' \succcurlyeq P \succ (P' \cup Q) \succcurlyeq (P \cup Q) \succ Q)$, provided at least one $\succcurlyeq$ can be replaced by $\succ$

In light of this principle, **MON4** effectively says that an agent's preferences should never be such that for any $P, P' \in \mathcal{P}$, it's not the case that both $P \succcurlyeq^{b} P'$ and $P' \succ^{b} P$.

The next few preference conditions require us to characterise a set of *null propositions*:

Definition 6.4: Null propositions
$\mathcal{N} = \{P \in \mathcal{P}\colon (P \cup Q) \sim Q$ for some $Q \in \mathcal{P}$ such that $(P \cap Q) = \emptyset$ and $\neg(P \sim Q)\}$

As with other definitions of nullity, the idea makes intuitive sense from a pragmatic standpoint. If the agent is not indifferent between some pair of disjoint propositions P and Q, then she should only be indifferent between $(P \cup Q)$ and Q if she is entirely confident that P is false. We can then assume, given her degrees of belief, that the news that $(P \cup Q)$ is true is in effect just the news that Q is true. This idea is formalised in **MON5**:

MON5 If $P \in \mathcal{N}$, then $(P \cup Q) \sim Q$ for all $Q \in \mathcal{P}$

The next three preference conditions ensure the existence of a countably additive probability function $\mathcal{B}el$:

MON6 (i) $(P \cap \neg P) \in \mathcal{N}$; (ii) if $P \in \mathcal{N}$ and $Q \in \mathcal{P}$, then $(P \cap Q) \in \mathcal{N}$; (iii) If $\{P_1, P_2, P_3, \ldots\}$ is a countable subset of $\mathcal{N}$, then the disjunction of $\{P_1, P_2, P_3, \ldots\}$ is also in $\mathcal{N}$

MON7 Any collection of pairwise incompatible propositions in $\mathcal{P} - \mathcal{N}$ is countable

MON8 Let $\{P_1, P_2, P_3, \ldots\}$ be a countable set of pairwise incompatible propositions in $\mathcal{P}$ whose disjunction is P; then (i) if $Q \succcurlyeq (P_1 \cup P_2 \cup \ldots \cup P_n)$ for all n, then $Q \succcurlyeq P$, and (ii) if $(P_1 \cup P_2 \cup \ldots \cup P_n) \succcurlyeq Q$ for all n, then $P \succcurlyeq Q$

MON6 is an obvious requirement if $\mathcal{B}el$ on $\mathcal{P}$ is to behave like a probability function. The Archimedean axiom **MON7** has the effect of ruling out infinitesimal probabilities, and **MON8** ensures that $\mathcal{B}el$ is countably additive.

The representation theorem can then be stated thus:

Theorem 6.3: Jeffrey-Bolker theorem
If **MON0–MON8** hold of $<\mathcal{W}, \mathcal{N}, \mathcal{P}, \succcurlyeq>$, then there exists at least one countably additive probability function $\mathcal{B}el$ on $\mathcal{P}$ and a real-valued function $\mathcal{D}es^*$ on the atomic elements w of $\mathcal{P}$, whose associated conditional expected utility $\mathcal{D}es$ on $\mathcal{P}$ is such that for all P, Q, $(P \cup \neg P) \in \mathcal{P}$,

(i) $\mathcal{D}es(P) = \sum_W \mathcal{B}el(w|P).\mathcal{D}es^*(w)$

(ii) $\mathcal{D}es(P \cup \neg P) = 0$

(iii) $\mathcal{D}es(P) \geq \mathcal{D}es(\mathrm{Q})$ iff $P \succcurlyeq Q$

Furthermore, the pair $<\mathcal{B}el, \mathcal{D}es>$ is unique up to a fractional linear transformation

Note that the representation involves two utility functions: $\mathcal{D}es^*$ is defined only on the atomic elements of $\mathcal{P}$, while $\mathcal{D}es$ is defined for all $P \in \mathcal{P}$ and characterised in terms of $\mathcal{B}el$ and $\mathcal{D}es^*$.

The uniqueness properties of this representational system are quite different than those we find in other CEU theorems. Neither $\mathcal{B}el$, $\mathcal{D}es^*$, nor $\mathcal{D}es$ are unique; instead, the pair

<$\mathcal{B}el$, $\mathcal{D}es$> is unique up to a fractional linear transformation. $\mathcal{D}es$ is normalised such that $\mathcal{D}es(R \cup \neg R) = 0$ and $\mathcal{D}es(R) = 1$, for some proposition R satisfying **MON1**. Let *inf* be the greatest lower bound of the values assigned by $\mathcal{D}es$, and let *sup* designate the least upper bound. Finally, let λ be a parameter falling between -1/*inf* and -1/*sup*. Then the fractional linear transformation <$\mathcal{B}el_\lambda$, $\mathcal{D}es_\lambda$> of <$\mathcal{B}el$, $\mathcal{D}es$> corresponding to λ is given by:

$$\mathcal{B}el_\lambda(P) = \mathcal{B}el(P).(1 + \lambda\mathcal{D}es(P))$$
$$\mathcal{D}es_\lambda(P) = \mathcal{D}es(P).((1 + \lambda) / (1 + \lambda\mathcal{D}es(P))$$

Interestingly, fractional linear transformations of a <$\mathcal{B}el$, $\mathcal{D}es$> pair can alter not only the *absolute* values that $\mathcal{B}el$ assigns to propositions, but also their *relative* values; i.e., different possible representations of exactly the same system of preferences will sometimes disagree regarding which of two propositions has should be assigned a higher credence. More generally, an agent's preferences on this kind of monoset framework do not typically determine a unique relative credence ordering $\succcurlyeq^b$ on $\boldsymbol{\mathcal{P}}$. Jeffrey suggested that it would be possible to pin down a unique $\mathcal{B}el$ if $\succcurlyeq^b$ were treated as a primitive relation on par with $\succcurlyeq$, with its own set of conditions (e.g., in his 1974, 1983). Joyce (1999, 138ff) proved that this is possible.

6.2.2 Critical discussion

Jeffrey's monoset theorem is one of the best-known amongst philosophers. Amongst other disciplines, however, the monoset framework is often considered problematic. As Fishburn puts it,

> Although well known in certain philosophical circles, Jeffrey's work is infrequently cited, and by implication not widely known, in other disciplines that share the legacy of preference and decision theory … A casual search of works on the foundations of decision and relational measurement in the fields of psychology, economics, statistics and management science indicates that if Jeffrey's work is mentioned at all, it is likely to be in reference to *The Logic of Decision*, and then only to note that it proposes a theory of decision that differs from traditional paradigms. (1994, 136)

There are, consequently, very few representation theorems based on an ontologically similar framework. Two recent exceptions to this trend can be found in (Bradley 1998, 2007) and (Ahn 2008), and as noted above, Armendt's (1986) theorem is ontologically similar to Jeffrey's system in that it takes preferences to be defined on a set of propositions and the possible roulette lotteries that may be formed thereupon. To keep the discussion brief, I will focus my criticisms on Theorem 6.3—the main points to be discussed apply equally to the other theorems just mentioned.

As far as characterisational representationism is concerned, Theorem 6.3 seems to take us several steps in the right direction. In particular, it neither appeals to act-functions nor lottery-functions—two bugbears which we have seen create problems for the multiset theorems considered so far. Furthermore, the domain of its preference relation is not limited to some obscure class of entities (such as infinite conjunctions of counterfactuals or lotteries upon lotteries), but instead seems capable of encapsulating *everything* towards which we could have mentalistic preferences. Of course, a mentalistic construal of $\succcurlyeq$ means that Theorem 6.3 fails to satisfy the naturalistic desideratum (**5**), but we have seen that the standard strategies for trying to formulate a theorem around the behavioural notion of preference lead to far worse concerns for characterisational representationism. Finally, Theorem 6.3's $\mathcal{Bel}$ and $\mathcal{Des}$ are defined on precisely the same domain, a feature not shared by any of the multiset theorems we have considered so far (desideratum (**2a**)).

Theorem 6.3 does, however, have some limitations; these I will note below, though first I want to briefly discuss one characteristic of Theorem 6.3 that I don't take to be especially problematic—in particular, the theorem's relatively weak uniqueness conditions. These are often cited as a cause for concern, as though characterisational representationism must be based upon a theorem which comes with (at least) the Standard Uniqueness Condition. But it's difficult to see why this should be so.

There are at least two (not mutually exclusive) strategies by which a proponent of characterisational representationism might attempt to deal with Theorem 6.3's weak uniqueness results. First, one can appeal to information which goes beyond agents' (actual or counterfactual) preferences. This further information can be used to narrow down the range of potential interpretations whenever a representation theorem does come with strong uniqueness conditions. For example, if the theorem's $\mathcal{Bel}$ function is non-unique, a principle like Charity might be used to constrain the set of available $\mathcal{Bel}$ representations down to uniqueness (§4.2). Second, where a theorem supplies us with a restricted set of possible $\mathcal{Bel}$ and $\mathcal{Des}$ representations, we might take *the entire set* as a model of the agent's credences and utilities. After all, Theorem 6.3 *does* carry the implication that there is a unique *set* of <$\mathcal{Bel}$, $\mathcal{Des}$> pairs (each related to the others by a fractional linear transformation) such each such pair jointly T-represents $\succcurlyeq$ on $\mathcal{P}$. Perhaps, then, that unique set—the '*representor*'—might be used to jointly represent the agent's credences and utilities: roughly, whatever is true of *every* $\mathcal{Bel}$ in the set is true of the agent's credence state (and likewise for their utilities). So, for instance, if every $\mathcal{Bel}$ in the representor always assigns a higher value to P than to Q, then the agent's credence in P is higher than her credence in Q. Something close to this suggestion was briefly discussed in §5.2.4, and the idea was raised by Jeffrey in his (1983).

Neither of these two strategies comes without cost, of course. If the former is adopted, then the characterisation of credences and utilities *must* appeal to information that goes beyond the agent's preferences; this may be considered too much for some die-hard advocates of a very strict form of preference functionalism. On the other hand, if the latter strategy is adopted, then the very intuitive picture of an agent as an expected utility maximiser must be sacrificed for a rather more complex model involving the interaction total credence and utility states modelled by *sets* of $\mathcal{B}el$ and $\mathcal{D}es$ functions. Nevertheless, neither of these costs seems like a strong enough reason to reject the possibility of basing characterisational representationism on something like Theorem 6.3.

If there is a serious problem with Theorem 6.3, it relates to whether its $\mathcal{B}el$ and $\mathcal{D}es$ functions (or sets thereof) can serve as accurate models of an ordinary agent's credences and utilities (desideratum (**2**)). For one thing, the $\mathcal{B}el$ associated with Theorem 6.3 is always a probability function, which puts limits on the kinds of credence states that it can represent—though some of the issues here depend on whether the set $\mathcal{W}$ is taken to be a set of *possible worlds*. If it is, then $\mathcal{B}el$ is limited to the representation of probabilistically coherent agents—and *ipso facto* incapable of representing the average person. The same is true of any representor set constructed solely from probability functions: for instance, every probability function $\mathcal{P}r$ built on a space of possible worlds will assign 0 to impossible propositions, 1 to necessary propositions, and satisfies the property that if $P \vdash Q$, then $\mathcal{P}r(P) \leq \mathcal{P}r(Q)$.

It may, however, be possible to avoid this implication by letting $\mathcal{W}$ be composed of both possible and impossible worlds (Nolan 1997), although taking this route may lead to other concerns (see, e.g., Bjerring 2013). Another problem, however—and one that an appeal to impossible worlds won't help with—is that if $\mathcal{B}el$ is to be a probability function, then its domain $\mathcal{P}$ must be closed under (at least finite) disjunctions, yet it may be too much to ask of ordinary agents that they have credences towards *every* disjunction $P \vee Q$ which can be formed from the propositions P and Q towards which they do have credences (desideratum (**2c**)). Worse still, in Jeffrey's system, $\mathcal{P} - \mathcal{N}$ is required to be a bottomless algebra, so $\mathcal{B}el$ and $\mathcal{D}es$ must be defined on a collection of ever-increasingly more specific propositions—propositions which quickly become far too specific for any ordinary agent to contemplate.[84] And finally, $\mathcal{D}es(P)$ always equals the $\mathcal{B}el$-weighted average utility of the various different ways that P might come true. It is implausible that ordinary agents' utilities are so consistently rational in this way.

I do not consider these problems to be especially devastating, at least if the task is to develop a version of characterisational representationism aimed *solely* at ideally rational agents. However, I think we can do better—in Chapter 8, I will develop a theorem which is ontologically very similar to Theorem 6.3, but with much less restricted $\mathcal{B}el$ and $\mathcal{D}es$ functions and a more plausible representation overall. Before getting to that, though, we

[84] Domotor (1978) proves a theorem similar to Bolker's for the case where $\mathcal{P}$ is finite. He relies, however, on a particularly strong condition that he calls *projectivity*, and his uniqueness condition is weaker than Theorem 6.3's.

need to conclude our review of the representation theorems currently on offer for characterisational representationism with the very first such theorem to have been developed: Frank Ramsey’s.

CHAPTER SEVEN

Ramsey and the Ethically Neutral Proposition

In his posthumously published 'Truth and Probability', Frank Ramsey sketches a proposal for the empirical measurement of credences, along with a corresponding set of conditions for a (somewhat incomplete) representation theorem intended to characterize the preference conditions under which this measurement process is applicable. Ramsey's formal approach is distinctive, deriving first a utility function to represent an agent's utilities, and then using this to construct their credence function. In specifying his measurement process and his conditions, however, Ramsey introduces the notion of an *ethically neutral proposition*, the assumed existence of which plays a key role throughout Ramsey's system.

The existence of such propositions has often been called into question. Ramsey's own definition of ethical neutrality presupposes the philosophically suspect theory of logical atomism. On other common ways of defining the notion, it's frequently noted that we lack good reasons for supposing that ethically neutral propositions exist, and in some cases we find that there are very good reasons for supposing that they cannot exist. Any system which relies on the existence of such propositions ought to be rejected.

In this chapter, I will first outline Ramsey's proposal in some detail (§7.1). This will help us to see why Ramsey thought he needed to introduce the notion of ethical neutrality, and why any theorem which appeals to ethically neutral propositions should be considered highly problematic (§7.2). In particular, I will argue that—whatever else may be the case—any system which requires ethically neutral propositions fails to satisfy desideratum **(1)**.

7.1 Ramsey's proposal

One of Ramsey's main goals in 'Truth and Probability' was to argue that the laws of probability supply for us the "logic of partial belief"" (1931, 166). His argument proceeds by first attempting to say what credences *are*, and on the basis of that understanding, showing that credences are probabilistically coherent.

Regarding the first step, of defining credences, Ramsey clearly had operationalist sympathies, asserting that the notion "has no precise meaning unless we specify more exactly how it is to be measured" (1931, 167). To be measured as having probabilistically coherent credences *is* (more or less), on this picture, to have probabilistically coherent credences, and anyone who can be measured through Ramsey's procedure at all will have credences conforming to the laws of probability. Note that the procedure was intended to be applicable to ordinary agents—Ramsey was not trying to define credences for some ideally rational being, but for the everyday person on the street (albeit not without some unavoidable idealisation).

Setting operationalism aside, it's easy to see in 'Truth and Probability' an early statement of something like preference functionalism: credences are to be understood through their role with respect to preferences when considered in conjunction with a total utility state. Ramsey writes that "the degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it" (1931, 169). Ramsey argues against characterising credences in terms of some introspectively accessible feeling had by a subject upon considering the relevant proposition. These arguments go well beyond operationalism, though I will not recapitulate them here. He concludes that "intensities of belief-feelings … are no doubt interesting, but … their practical interest is entirely due to their position as the hypothetical causes of beliefs *qua* bases of action" (1931, 172). On this more charitable interpretation, Ramsey advocates an early version of characterisational representationism, and his representation theorem can be seen as spelling out precisely the relevant functional roles associated with credence states.

In any case, Ramsey proposes to take as the theoretical basis of his measurement system a particular theory of decision making—that is, the theory that "we act in the way that we think most likely to realize the objects of our desires, so that a person's actions are completely determined by his desires and opinions" (1931, 173). As noted in §3.1, his idea was to assume the basic truth of something like classical expected utility theory, and on that assumption, use empirical information about an agent's preferences to work out what her credences and utilities must be. Ramsey was entirely aware of the empirical difficulties facing that theory, writing that:

> [it] is now universally discarded, but nevertheless comes, I think, fairly close to the truth in the sort of cases with which we are most concerned … This theory cannot be made adequate to all the facts, but it seems to me a useful approximation to the truth particularly in the case of our self-conscious or professional life, and it is presupposed in a great deal of our thought. (1931, 173)

We will return shortly to what Ramsey meant by "the sort of cases with which we are most concerned", and exactly what he needed to assume to get his measurement process off the ground.

We can summarise Ramsey's measurement procedure as follows:

(1) Determine S's preferences over worlds and gambles
(2) Define a relation of equal difference in utilities
(3) Locate ethically neutral propositions of credence ½
(4) Construct an interval scale representation $\mathcal{D}es$ of S's preferences
(5) Use $\mathcal{D}es$ to define a probability function $\mathcal{B}el$

I will discuss each step in turn. For the sake of simplicity, I have neglected to discuss one important aspect of Ramsey's procedure: the use of preferences over complex gambles to define conditional probabilities, which are used to show that the measured credences constitute a probability function.[85]

7.1.1 Determining a preference ordering

The first stage of Ramsey's procedure is to determine the agent's preferences over different ways the world might be. This is, according to Ramsey, relatively straightforward:

> If … we had the power of the Almighty, and could persuade our subject of our power, we could, by offering him options, discover how he placed in order of merit all possible courses of the world. In this way all possible worlds would be put in an order of value … (1931, 176)

Ramsey writes that he intends the relevant objects of preference to be "different possible totalities of events … the ultimate organic unities" (1931, 177-8)—that is, possible worlds. I will use $\mathcal{O} = \{o_1, o_2, \ldots\}$ to designate the set of these "possible totalities of events", which I'll refer to as *outcomes*. Importantly, however, within only a few paragraphs, Ramsey goes on to note that with respect to *at least one* proposition P, and *some* o_1, o_2, '[o_1] and [o_2] must be supposed so far undefined as to be compatible with both P and $\neg P$' (1931, 178). The most natural interpretation of this seems to be that in some select few circumstances o_1 and o_2 ought to be considered *not quite* as worlds, but rather as propositions maximally specific with respect to everything *except* P.

I suspect that Ramsey would have been happy with letting $\mathcal{O}$ be a set of consistent propositions which are only maximally specific with respect to what the agent cares about, and making this exegetical move resolves certain difficulties which appear elsewhere in his theory (see §7.2.1). However, in what follows we will simply treat $\mathcal{O}$ as a set of very highly specific consistent propositions, some of which—but *not all*—may perhaps be maximally specific.

Given a preference ordering over $\mathcal{O}$, we are required then to empirically determine how the agent ranks *gambles*. Once again, Ramsey asks us to imagine that we have "persuaded

[85] This part of Ramsey's procedure is outlined in (Bradley 2001).

our subject of our power”, but this time we make offers of the following kind: “Would you rather have world [o_3] in any event, or world [o_1] if P is true, and world [o_2] if P is false?” (1931, 177). Let us represent the latter option, the gamble o_1 *if P is true*, o_2 *otherwise*, as simply (o_1, P; o_2). Ramsey then notes that:

> If … [the agent] were certain that P was true, he would simply compare [o_1] and [o_3] and choose between them as if no conditions were attached; but if he were doubtful his choice would not be decided so simply. (1931, 177)

Here, Ramsey looks to compare an outcome with a gamble, so we are to assume that gambles and outcomes are comparable. It is also evident from the conditions he later provides that we need to consider agents’ preferences between gambles. In sum, if $\mathcal{G}$ is the set of all gambles of the form (o_1, P; o_2), then Ramsey requires us to empirically determine a preference ordering on $\mathcal{O} \cup \mathcal{G}$.

There are a number of interpretive difficulties with Ramsey’s proposal that might be raised at this point. Contrary to what is frequently claimed, Ramsey nowhere mentions preferences over *acts*—and indeed, his simple two-outcome gambles lack sufficient structure to plausibly represent any act (whether objectively or as the agent in question conceives of them). It is unclear, however, how Ramsey intended for us to understand his gambles. For reasons outlined by Joyce (1999, 62-3), “o_1 if P is true, and o_2 if P is false” should not be understood using material conditionals. Sobel (1998, 239) suggests that (o_1, P; o_2) is just a conjunction of subjunctives,

$$(P \mathrel{\Box\!\!\rightarrow} o_1) \mathbin{\&} (\neg P \mathrel{\Box\!\!\rightarrow} o_2)$$

On the other hand, Bradley (1998, 193-4) treats his Ramsey-style gambles as a conjunction of indicative conditionals. Because he also wants to accept Adams’ Thesis (see Adams 1975), he foregoes any propositional interpretation of his theory’s (o_1, P; o_2), and instead treats his analogue of $\mathcal{G}$ as a set of sentences in a formal language. I will not attempt to adjudicate whether it’s better to use subjunctive or indicative conditionals here—the issues that I will discuss are independent of any concerns that one might have here.

It would be a mistake—though one which is unfortunately common—to equate a disposition to *choose* one gamble over another with a preference for the truth of one conjunction of (indicative or subjunctive) conditionals over another. Dispositions to choose between gambles will depend on how the agent in question conceives of the options available, and there is no guarantee that by offering S a gamble Γ that returns o_1 if P is true, o_2 otherwise, S will thereby *represent* Γ as such—S may have misheard, or may not trust the offer.[86] A rather more plausible claim, then, would be to say that a disposition to choose one gamble over another *goes hand in hand with* a preference for the truth of one pair of conditionals over another—*viz*., that pair of conditionals which the agent believes would *most likely* be made true by her choice.

It is also unclear how propositions as highly specific as Ramsey suggests can be offered to any ordinary human subject; the power to conceptualise even one possible world in all its detail seems beyond the average person. Even more worrying is that convincing a subject that "we had the power of the Almighty" would surely drastically alter her doxastic state prior to measuring it, as Jeffrey (1983, 158-60) has noted. Likewise, when a subject is offered the choice of either o_3 or $(o_1, P; o_2)$, we must not suppose that her credence in P is in any way changed by the offer, or this would ruin the measurement.

Interestingly, Ramsey himself objects to the betting interpretation of credences on the grounds that "the proposal of the bet may inevitably alter [the subject's] state of opinion" (1931, 172). Either Ramsey did not recognise that the same objection applies with greater force to his own account, or he believed that the worry could be addressed. Bradley (2001, 285-8) suggests one way in which it might be addressed: rather than making the subject believe in our godlike powers, we simply ask her to judge her preferences amongst options *as if* they were genuinely available to her. To the extent that such a request can be satisfied, this re-construal of Ramsey's methodology may help to minimise any changes to subjects' credences prior to measurement.

In any case, we can now say precisely what Ramsey meant when he referred to the accuracy of expected utility theory in "the sort of cases with which we are most concerned". We are to limit our attention to conscious, deliberate and presumably reflective judgements of preference between outcomes and outcomes, gambles and gambles, and outcome and gambles. Plausibly, Ramsey would have also held that we are not to consider cases where the subject is intoxicated, or under any kind of substantial physical or emotional duress. Ramsey does not need to assume anything as strong as the truth of classical expected utility theory *tout court*, nor even its approximate truth across a wide range of cases—he only needs that it is accurate in this particular kind of case.

Although his own use of $\succcurlyeq$ is generally put in behavioural terms, I do not think that it would be very harmful to the essence of Ramsey's account to interpret his $\succcurlyeq$ as a kind of *considered mentalistic* preference relation; roughly:

> $x \succcurlyeq y$ relative to an agent S iff S judges x to be at least as good as y after consciously deliberating on the matter, while neither under physical or emotional distress, nor under the influence of any intoxicating substances, and so on

[86] The point here is similar, of course, to the one raised in §5.4 regarding the interpretation of Savage's act-functions. See Chapter 9 for further discussion.

On this interpretation, Ramsey's assertion that expected utility theory is broadly accurate in "the sort of cases with which we are most concerned" is essentially the claim that an ordinary agent's reflective preferences are what we would expect of an expected utility maximiser.

7.1.2 Defining an equal difference relation

Ramsey's first step has us empirically determine how the agent ranks outcomes and gambles. However, a simple preference ordering on outcomes and gambles only suffices for an ordinal scale representation of an agent's utilities for those outcomes and gambles. For Ramsey, this is unsatisfactory: "There would be no meaning in the assertion that the difference in value between [o_1] and [o_2] was equal to that between [o_3] and [o_4]" (1931, 176). Thus Ramsey sets himself the task of characterizing an *equal difference* (in utilities) relation between pairs of outcomes wholly in terms of preferences over gambles. If he can do this, then on the basis of well-known results from the mathematical theory of measurement, he can construct a richer representation of our utilities.

Let us say that $(o_1, o_2) =^d (o_3, o_4)$ holds iff the difference in value for the agent between o_1 and o_2 is equal to the difference in value between o_3 and o_4. Ramsey's goal of defining $=^d$ in terms of preferences over gambles then sets up a certain difficulty to be overcome. According to the background assumption of CEU, an agent's preferences over gambles are determined by two factors: their *utilities* and their *credences*. Whether an agent prefers $(o_1, P; o_2)$ to $(o_3, Q; o_4)$, for example, depends partly on the utilities that she attaches to o_1, o_2, o_3, o_4, and partly on the credences regarding P and Q. However, whether $(o_1, o_2) =^d (o_3, o_4)$ holds for that agent should depend *solely* on the utilities she attaches to o_1, o_2, o_3, o_4. In order to define $=^d$ in terms of preferences over gambles, then, Ramsey needs some way of factoring out any confounding influences, so that whether the agent prefers one of the relevant gambles to another depends *only* on the utilities attached to the outcomes involved.

Ramsey's central innovation here is to define, in terms of preference, what it is for an agent to have credence ½ in a proposition, and then to use this to define $=^d$. Let us suppose for now that whether an agent prefers $(o_1, P; o_2)$ to $(o_3, Q; o_4)$ depends *only* on the utilities the agent has for o_1, o_2, o_3, o_4, and the credences she has for P and Q. More specifically, assume *Naïve Expected Utility Theory*:

> **Naïve Expected Utility Theory**
> If $\mathcal{D}es$ is a real-valued function that models the agent's utilities, and $\mathcal{B}el$ is a credence function that models the agent's credences, then $(o_1, P; o_2) \succcurlyeq (o_3, Q; o_4)$ iff $\mathcal{D}es(o_1).\mathcal{B}el(P) + \mathcal{D}es(o_2).(1 - \mathcal{B}el(P)) \geq \mathcal{D}es(o_3).\mathcal{B}el(Q) + \mathcal{D}es(o_4).(1 - \mathcal{B}el(Q))$

We will note shortly that Ramsey did *not* assume Naïve Expected Utility Theory; but for now it suffices to explain the reasoning behind his definitions. It is worth noting that while $\mathcal{B}el$ is only required to be a *credence function* (rather than a probability function specifically), Naïve Expected Utility Theory does carry the implicit assumption that $\mathcal{B}el(\neg P) = 1 - \mathcal{B}el(P)$. Were this *not* the case, we would expect the contribution of o_2 to the desirability of $(o_1, P; o_2)$ to be determined by $\mathcal{D}es(o_2).(1 - \mathcal{B}el(P))$ rather than $\mathcal{D}es(o_2).\mathcal{B}el(\neg P)$ directly.

Suppose that the agent is indifferent between $(o_1, P; o_2)$ and $(o_2, P; o_1)$. According to Naïve Expected Utility Theory, there are only two (not mutually exclusive) ways in which this might come about: either both o_1 and o_2 have exactly the same utility for the agent, or the agent's credence in P is exactly ½. To rule out the former possibility, we consider a pair of gambles $(o_3, P; o_4)$ and $(o_4, P; o_3)$, where we know that the agent is not indifferent between o_3 and o_4. If we find that the agent is indifferent between $(o_3, P; o_4)$ and $(o_4, P; o_3)$, we will have established that $\mathcal{B}el(P) = ½$. If her credence in P were any other way, then the agent would have not been indifferent between $(o_3, P; o_4)$ and $(o_4, P; o_3)$.

With this in place, we are then able to say that $(o_1, o_2) =^d (o_3, o_4)$ holds iff $(o_1, P; o_4) \sim (o_2, P; o_3)$, where P is such that the agent believes it to degree ½. The reasoning behind this is not immediately obvious. From the assumption of Naïve Expected Utility Theory, we have that $(o_1, P; o_4) \sim (o_2, P; o_3)$ holds just in case:

$$\mathcal{D}es(o_1).\mathcal{B}el(P) + \mathcal{D}es(o_4).(1 - \mathcal{B}el(P)) = \mathcal{D}es(o_2).\mathcal{B}el(P) + \mathcal{D}es(o_3).(1 - \mathcal{B}el(P))$$

We have also already established that $\mathcal{B}el(P) = ½ = 1 - \mathcal{B}el(P)$, so we can drop the constant factor leaving us with:

$$\mathcal{D}es(o_1) + \mathcal{D}es(o_4) = \mathcal{D}es(o_2) + \mathcal{D}es(o_3)$$

Which holds just in case:

$$\mathcal{D}es(o_1) - \mathcal{D}es(o_2) = \mathcal{D}es(o_3) - \mathcal{D}es(o_4)$$

This just states that the difference between o_1 and o_2 is equal to the difference between o_3 and o_4; so if $\mathcal{B}el(P) = ½$, $(o_1, P; o_4) \sim (o_2, P; o_3)$ iff $(o_1, o_2) =^d (o_3, o_4)$.

7.1.3 Locating ethically neutral propositions

Before moving on to measuring utilities, however, Ramsey makes the following note:

> There is first a difficulty which must be dealt with; the propositions like P … which are used as conditions in the [gambles] offered may be such that their truth or falsity is an object of desire to the subject. This will be found to complicate the problem, and we have to assume that there are propositions for which this is not the case, which we shall call ethically neutral. (1931, 177)

This is the entirety of what Ramsey writes regarding his motivation for introducing ethically neutral propositions.

The idea is clear enough: Naïve Expected Utility Theory is mistaken, as it fails to take into account the utility that may attach to the gamble's condition and how the condition might influence the agent's valuation of the elements of $\mathcal{O}$. Assuming that o_1 is consistent with both P and $\neg P$, it's possible that an agent might attach a different value to (o_1 & P) than to (o_1 & $\neg P$). These are potentially quite different states of affairs with potentially different utilities, and the truth or falsity of P might make a great deal of difference to how the outcome o_1 is valued. For instance, suppose that in o_1 the agent has a puppy as a pet, while in o_2 she instead keeps a kitten, and let P be *puppies spread disease but kittens don't*; plausibly, (o_1 & P) will be valued quite differently than (o_1 & $\neg P$), and likewise for (o_2 & P) and (o_2 & $\neg P$).

Instead of Naïve Expected Utility Theory, and supposing o_1, o_2, o_3, and o_4 are each compatible with the relevant propositions, we should really have that:

$$(o_1, P; o_2) \succcurlyeq (o_3, Q; o_4)$$

Just in case:

$$\mathcal{Des}(o_1 \,\&\, P).\mathcal{Bel}(P) + \mathcal{Des}(o_2 \,\&\, \neg P).(1 - \mathcal{Bel}(P)) \geq \mathcal{Des}(o_3 \,\&\, Q).\mathcal{Bel}(P) + \mathcal{Des}(o_4 \,\&\, \neg Q).(1 - \mathcal{Bel}(Q))$$

It is easy to see that this fact invalidates the reasoning behind both the definition of what it is for an agent to have a credence ½ in a proposition, and the definition of $=^{d}$, for now we can no longer say that the agent's preferences between (o_1, P; o_2) and (o_3, Q; o_4) depend on their credences in P and Q and the utilities the agent has for o_1, o_2, o_3, o_4. Rather, they actually depend on the agent's credences in P and Q and utilities for (o_1 & P), (o_2 & $\neg P$), (o_3 & Q), and (o_4 & $\neg Q$).

Ramsey's solution to this difficulty is the ethically neutral proposition—a kind of proposition the truth or falsity of which is of absolutely no concern to the agent. Ramsey provides us with a problematic definition of the notion, which I will discuss further in §7.2.2. The apparent purpose of its introduction, however, is that if P is ethically neutral, then the conjunction of P with o has the same utility as o itself, and similarly for the conjunction of $\neg P$ and o. Setting aside Ramsey's own definition, we can say that P is *ethically neutral* whenever $o \sim (o \,\&\, P) \sim (o \,\&\, \neg P)$, for any $o \in \mathcal{O}$ that is compatible with both P and $\neg P$.

So long as we are considering gambles conditional on ethically neutral propositions, we can *without risk of error* apply Naïve Expected Utility Theory. Thus Ramsey happens upon the following two definitions:

> **Definition 7.1: Ethically neutral proposition of credence ½**
> P is an ethically neutral proposition of credence ½ iff P is ethically neutral, and for some o, $o_2 \in \mathcal{O}$, $\neg(o_1 \sim o_2)$, and $(o_1, P; o_2) \sim (o_2, P; o_1)$

And:

> **Definition 7.2: Equal difference relation**
> $(o_1, o_2) =^d (o_3, o_4)$ iff $(o_1, P; o_4) \sim (o_2, P; o_3)$, where P is an ethically neutral proposition of credence ½

7.1.4 Measuring utilities

At this point, Ramsey lists eight preference conditions, and states (but does not prove) that their satisfaction enables an appropriately rich representation of the agent's preferences. Let $\mathcal{P}$ be a set of propositions, $\mathcal{O}$ the set of outcomes, and $\mathcal{G}$ the set of gambles; $\succcurlyeq$ is defined on $\mathcal{O} \cup \mathcal{G}$. Ramsey's Representation Conjecture can then be stated thus:

> **Ramsey's Representation Conjecture**
> If **RAM1–8** hold of $<\mathcal{P}, \mathcal{O}, \mathcal{G}, \succcurlyeq>$, then there exists a real-valued function $\mathcal{D}es$ on $\mathcal{O}$ such that for all $o_1, o_2, o_3, o_4 \in \mathcal{O}$,
>
> (i) $\mathcal{D}es(o_1) - \mathcal{D}es(o_2) = \mathcal{D}es(o_3) - \mathcal{D}es(o_4)$ iff $(o_1, o_2) =^d (o_3, o_4)$
>
> Furthermore, $\mathcal{D}es$ is unique up to positive linear transformation

We will not consider whether Ramsey's preference conditions successfully ensure the desired representation result, or how they might be fleshed out to do so if not—though see (Bradley 2001) for relevant work in this regard. It is clear that something in the vicinity of Ramsey's conditions should suffice, though I will not take a stand on the precise formulation needed.

The very first preference condition is the most distinctive aspect of Ramsey's theorem:

> **RAM1** There is at least one ethically neutral proposition of credence ½

The importance of **RAM1** for the rest of Ramsey's formal system should not be understated. Most of the preference conditions to follow are stated in terms of $=^{\mathrm{d}}$, which is defined in terms of ethically neutral propositions. If **RAM1** is false, those conditions will be in some cases false, in others trivial; in either case, the system as a whole collapses without this foundational assumption.

The next three preference conditions are each obviously necessary for Ramsey's desired representation result. For all P, $Q \in \mathcal{P}$, o_1, o_2, o_3, o_4, o_5, $o_6 \in \mathcal{O}$, $(o_1, P; o_2)$, $(o_3, P; o_4) \in \mathcal{G}$, and $x, y, z \in \mathcal{O} \cup \mathcal{G}$,

RAM2 (i) If P, Q, are both ethically neutral propositions of credence ½, and $(o_1, P; o_2) \sim (o_3, P; o_4)$, then $(o_1, Q; o_2) \sim (o_3, Q; o_4)$, and (ii) if $(o_1, o_2) =^{\mathrm{d}} (o_3, o_4)$, then $o_1 \succ o_2$ iff $o_3 \succ o_4$, and $o_1 \sim o_2$ iff $o_3 \sim o_4$

RAM3 $\sim$ is transitive

RAM4 $=^{\mathrm{d}}$ is transitive

The role of **RAM2** is ensure that the definition of $=^{\mathrm{d}}$ is coherent. Together, **RAM2**–**RAM4** help to ensure that $=^{\mathrm{d}}$, which holds between pairs of outcomes, mirrors the behaviour of the *equals* relation between the differences of pairs of real numbers.

The following two existential conditions are stated in terms of what Ramsey calls *values*. Formally,

Definition 7.3: The value of *o*

For every $o \in \mathcal{O}$, let $\underline{\underline{o}} = \{o' \in \mathcal{O}: o' \sim o\}$

The value of an outcome o, denoted $\underline{\underline{o}}$, is the set of all outcomes in $\mathcal{O}$ with the same desirability as o. Ramsey's next two conditions are then:

RAM5 For all $\underline{\underline{o}}_1$, $\underline{\underline{o}}_2$, $\underline{\underline{o}}_3$, there is exactly one $\underline{\underline{o}}_4$ such that $(o_1, o_4) =^{\mathrm{d}} (o_2, o_3)$

RAM6 For all $\underline{\underline{o}}_1$, $\underline{\underline{o}}_2$, there is exactly one $\underline{\underline{o}}_3$ such that $(o_1, o_3) =^{\mathrm{d}} (o_3, o_2)$

RAM5 implies that there is always at least one outcome o_4 such that the difference between o_1 and o_4 is equal to the difference between o_2 and o_3, for *any* choice of outcomes o_1, o_2 and o_3. In a manner of speaking, **RAM6** says that for any pair of outcomes o_1 and o_2, there is at least one outcome o_3 with a utility exactly half-way between that of o_1 and o_2. Given **RAM1** (which implies the non-triviality of $\succ$ on $\mathcal{O}$), this entails a *denseness* to the agent's preference structure—and correspondingly, that $\mathcal{O}$ is infinite.

Finally, Ramsey lists two other conditions, which are not spelled out in any detail:

RAM7 “Axiom of continuity:—Any progression has a limit (ordinal)” (Ramsey 1931, 179)

RAM8 Archimedean condition

What Ramsey intended for **RAM7** is something of a mystery. One guess (cf. Sobel 1998, Bradley 2001) would be that for every gamble (o_1, P; o_2), there is an outcome o_3 such that $o_3 \sim (o_1, P; o_2)$. A condition to this effect seems to be required to ensure that every real number can be mapped to at least one outcome’s value.

Ramsey does not specify the character of **RAM8**, however it’s easy to guess its role—like other so-called Archimedean conditions in various representation theorems, it is supposed to rule out any one outcome or gamble being incomparably better or worse than another. More specifically, **RAM8** ensures that the numerical representation satisfies the Archimedean property of real numbers: for any positive number x, and any number y, there is an integer n such that $n + x \geq y$.[87]

7.1.5 Measuring credences

Suppose that we have our function $\mathcal{D}es$. Ramsey then argues that:

> Having thus defined a way of measuring value we can now derive a way of measuring belief in general. If the option of [o_2] for certain is indifferent with that of [(o_1, P; o_3)], we can define the subject’s degree of belief in P as the ratio of the difference between [o_2] and [o_3] to that between [o_1] and [o_3] … This amounts roughly to defining the degree of belief in P by the odds at which the subject would bet on P, the bet being conducted in terms of differences of value as defined. (1931, 179-80)

In a footnote, Ramsey adds that ‘[o_1] must include the truth of P, [o_3] its falsity; P need no longer be ethically neutral’ (1931, 179). We are led to the following definition:

Definition 7.4: Ramsey’s *Bel*
For all contingent propositions P and outcomes o_1, o_2, o_3 such that o_1 implies P, o_3 implies $\neg P$, $\neg(o \sim o_3)$, and $o_2 \sim (o_1, P; o_3)$, $\mathcal{B}el(P) = (\mathcal{D}es(o_2) - \mathcal{D}es(o_3))/(\mathcal{D}es(o_1) - \mathcal{D}es(o_3))$

Ramsey mistakenly states that Definition 7.4 “only applies to partial belief and does not include certain beliefs” (1931, 180), though perhaps he meant that the definition does not

[87] Were one to spell out **RAM8**, it is likely that it would need to look much like **ADS5** of Definition 8.6 below.

apply if P is *non-contingent*. In this case, we simply stipulate that $\mathcal{B}el(P) = 1$ if P is necessary, 0 if P is impossible. Note that, because ratios of differences are preserved across positive linear transformations of $\mathcal{D}es$, $\mathcal{B}el(P)$ so-defined is unique.

The reasoning behind this final step is again left up to the reader, though also it follows from his background assumption of the descriptive adequacy of classical expected utility theory. Note first of all that if o_1 entails P, then the conjunction of P and o_1 is equivalent to o_1, so (Ramsey implicitly assumes) $\mathcal{D}es(o_1) = \mathcal{D}es(o_1 \,\&\, P)$. Thus, if $(o_1, P; o_2) \sim o_3$, where o_1 entails P and o_2 entails $\neg P$, then:

$$\mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es(o_1).\mathcal{B}el(P) + \mathcal{D}es(o_2).(1 - \mathcal{B}el(P)) = \mathcal{D}es(o_3)$$

This is then rearranged to give us the definition of $\mathcal{B}el(P)$ as above.

For future discussion, it is worth making Ramsey's implicit assumption explicit:

Indifference to Equivalent Conjunctions
For all P, Q, if $P \vdash Q$, then $P \sim (P \,\&\, Q)$

Ramsey does note two more assumptions needed to ensure the coherence of his definition. The first of these is that the value of $\mathcal{B}el(P)$ does not depend on the choice of outcomes and gambles satisfying the stated conditions. In effect, this is to place restrictions directly upon $\mathcal{B}el$ after it has been defined in terms of preferences. The second assumption is that for any gamble $(o_1, P; o_2)$ we will always be able to find some outcome o_3 such that $o_3 \sim (o_1, P; o_2)$.

Ramsey (1931, 180ff) goes on to define conditional probabilities using preferences over more complicated gambles, and he argues that $\mathcal{B}el$ satisfies the laws of probability, though I will not recapitulate that argument here: it is enough that Ramsey provides a credence function, $\mathcal{B}el\colon \boldsymbol{\mathcal{P}} \mapsto [0, 1]$, that supposedly represents the agent's credences—after all, it combines with the agent's utilities for outcomes to determine their preference ordering for two-outcome gambles in more or less the manner we pre-theoretically expect credence to do so. For our present purposes, it is incidental whether $\mathcal{B}el$ satisfies the conditions of the probability calculus.

7.2 The problem of ethical neutrality

Despite its very early inception, there are several features that make Ramsey's system attractive, especially in comparison to later works. The theorems developed by von Neumann and Morgenstern (1944) and Anscombe and Aumann (1963) were in some respects a rediscovery of ideas already present in 'Truth and Probability', but their appeal to extrinsically given probabilities limits their applicability, whereas Ramsey's system makes no such appeal. Savage's theorem was also founded on Ramseyan ideas, but Savage's system suffers from a number of defects not present in Ramsey's system. For instance, given the plausible assumption that Ramsey wanted to avoid impossible gambles (§7.2.1), the outcomes of a gamble are always consistent with the gamble's condition. Consequently, Ramsey's system seems to avoid anything like the constant acts problem that plagues Savage's system. Furthermore, the domain of Ramsey's $\mathcal{B}el$ is not limited to disjunctions of states. Another attractive feature of Ramsey's proposal is that it provides us with the Standard Uniqueness Condition. We might contrast this with the monoset theorem of §6.2, where the $<\mathcal{B}el, \mathcal{D}es>$ pair is only unique up to a fractional linear transformation.

All of this is achieved, however, on the basis of a highly problematic assumption about ethically neutral propositions, which I will now argue makes Ramsey's system untenable. My critical discussion of Ramsey's ideas focuses on this assumption as it raises unique problems not faced by the theorems I have considered in earlier chapters.

7.2.1 Why Ramsey needed ethical neutrality

Ramsey was right to reject Naïve Expected Utility Theory. If o is compatible with both P and $\neg P$, then it's entirely possible that the agent values (o & P) more (or less) than (o & $\neg P$). Any rational agent ought to take this into account when deliberating between gambles conditional on P with o as an outcome. For example, contrary to Naïve Expected Utility Theory, it's possible that the agent could be indifferent between $o_1 \sim o_2$ without thereby being indifferent between $(o_1, P; o_2)$ and $(o_2, P; o_1)$, if the truth or falsity of P makes a difference to how the agent values o_1 or o_2.

However, this point is conditional on o_1 and o_2 being each compatible with both P and $\neg P$. If instead we suppose that o_1 implies P, then (o_1 & P) is logically equivalent to o_1—and for Ramsey, if o_1 implies P, then the desirability of o_1 is just the desirability of (o_1 & P). Ramsey's characterisation of the $\mathcal{B}el$ function relies on this assumption. So, inasmuch as o_1 implies P and o_2 implies $\neg P$,

$$\begin{aligned}\mathcal{D}es((o_1, P; o_2)) &= \mathcal{D}es(o_1 \,\&\, P).\mathcal{B}el(P) + \mathcal{D}es(o_2 \,\&\, \neg P).(1 - \mathcal{B}el(P)) \\ &= \mathcal{D}es(o_1).\mathcal{B}el(P) + \mathcal{D}es(o_2).(1 - \mathcal{B}el(P))\end{aligned}$$

Note that this holds regardless of whether P is ethically neutral or not. In other words, if o_1 implies P and o_2 implies $\neg P$, and given Indifference to Equivalent Conjunctions, we can apply Naïve Expected Utility Theory to the gamble $(o_1, P; o_2)$.

Interestingly, Ramsey originally describes his outcome set $\mathcal{O}$ as a set of possible worlds, and it is part of Ramsey's background theory that every world individually determines the truth or falsity of any proposition. In particular, Ramsey assumed a broadly

Wittgensteinian logical atomism—though he believed it possible to reformulate his theorem without these commitments (see his 1931, 177). We are to suppose that there exists a class of atomic propositions such that no two worlds are exactly identical with respect to the truth of these propositions, every atomic proposition can be true or false entirely independently of any others, and for every world w and atomic proposition P, there is another world w^* that differs only with respect to the truth of P. Every possible world on this picture is determined by the set of atomic propositions true at that world. Even setting aside the assumption of logical atomism, on an orthodox conception of propositions as sets of worlds, then for any given (determinate) proposition, a given world either is or is not a member of that proposition. Every world therefore determines either the truth or falsity of any proposition.

This leaves us with something of a puzzle: why did Ramsey alter his characterisation of the outcome set (as noted in §7.1.1)? It seems that if he limited his attention to gambles like $(o_1, P; o_2)$, where o_1 implies P and o_2 implies $\neg P$, then he could have used preferences over *these* to define $=^d$ without needing to introduce the notion of ethical neutrality. The following piece of terminology will be helpful:

Definition 7.5: Impossible gambles
A gamble $(o_1, P; o_2)$ is *impossible* iff P and $\neg P$ are consistent and either $(o_1$ & $P)$ or $(o_2$ & $\neg P)$ are inconsistent; $(o_1, P; o_2)$ is *possible* otherwise

Where outcomes are possible worlds, every *possible* gamble $(o_1, P; o_2)$ conditional on a contingent proposition P must be such that o_1 implies P and o_2 implies $\neg P$. Where one of either P or $\neg P$ is impossible—say, $\neg P$—then the other must be necessary; in which case $\mathcal{B}el(\neg P) = 0$, $\mathcal{B}el(P) = 1$, and every o implies P, so $\mathcal{D}es(o) = \mathcal{D}es(o \,\&\, P)$. We can therefore *always* apply Naïve Expected Utility Theory to *possible* gambles, *if* the outcomes in $\mathcal{O}$ are worlds. So why did Ramsey not stick to his original characterisation of outcomes as worlds, and simply use preferences over possible gambles to define $=^d$?

The answer to this question can be discovered by considering again how Ramsey defines what it is for an agent to have a credence of ½ in a proposition. In particular, to determine whether P is of credence ½, we need to consider preferences over two gambles of the form $(o_1, P; o_2)$ and $(o_2, P; o_1)$. The definition Ramsey gives us *only* makes sense if the outcomes o_1 and o_2 are *not* possible worlds. If o_1 and o_2 are possible worlds, then at least one of the two gambles is impossible, and if either gamble is impossible, then the reasoning behind the assignment of a credence value of ½ to the contingent proposition P is no longer valid.

Indeed, Ramsey recognised the difficulty here, and for this reason wrote that, at least for some outcomes o_1 and o_2 required for his definition, o_1 and o_2 "must be supposed so far undefined as to be compatible with both P and $\neg P$". Supposing for simplicity that P is atomic, we are presumably to take the outcomes o_1 and o_2 as *near-worlds*, which we can understand as propositions that are just shy of being maximally specific. Given his logical atomism, for every world w and every atomic proposition P, there is a proposition that *nearly* uniquely identifies w except for specifying whether P is true or not. In Ramsey's framework, a near-world with respect to an atomic proposition P is a disjunction of two worlds w^P and $w^{\neg P}$ that are identical with respect to all of their atomic propositions except for P.

The answer to our puzzle, then, is that Ramsey's set of outcomes cannot quite be the set of possible worlds *given* his strategy for defining $=^{\mathrm{d}}$. For the pair of possible gambles $(o_1, P; o_2)$ and $(o_2, P; o_1)$ referred to in Definition 7.1, neither o_1 nor o_2 can imply either P or $\neg P$. It follows for the reasons given, then, that we cannot in general apply Naïve Expected Utility Theory to such gambles without appeal to ethically neutral propositions.

Before I move on to the issues surrounding ethically neutral propositions, it is worth noting that Ramsey's **RAM1** seems to *understate* what he actually required. This is because, given how he proposed to define $=^{\mathrm{d}}$, without changes elsewhere in his system Ramsey also required either that we have preferences over impossible gambles, or that *every* outcome in $\mathcal{O}$ was compatible with both the truth and falsity of *some* ethically neutral proposition. Suppose that $o_1 \sim o'_1$, so $\mathcal{D}es(o_1) - \mathcal{D}es(o'_1) = \mathcal{D}es(o'_1) - \mathcal{D}es(o_1)$. From Definition 7.2, we know that $(o_1, o'_1) =^{\mathrm{d}} (o'_1, o_1)$ is only defined if the agent has preferences over some pair of gambles of the form $(o_1, P; o_1)$ and $(o'_1, P; o'_1)$, where P is a contingent proposition. It follows that either o_1 is compatible with P and $\neg P$, and similarly for o'_1, or at least one of these two gambles is impossible.

One might suppose that Ramsey was happy to deal with preferences over impossible gambles. This would have forced him to assume that there is an interesting difference between two impossible propositions $(o_1 \,\&\, P)$ and $(o_2 \,\&\, P)$, where both o_1 and o_2 entail $\neg P$ but $\neg(o_1 \sim o_2)$. For suppose that Ramsey had only one impossible proposition, $\bot$. Then $\mathcal{D}es(o_1 \,\&\, P) = \mathcal{D}es(o_2 \,\&\, P) = \mathcal{D}es(\bot)$, but $\mathcal{D}es(o_1) \neq \mathcal{D}es(o_2)$. For whatever value we take $\mathcal{D}es(\bot)$ to have, it is clear that this will lead to problems. Suppose that $\mathcal{D}es(\bot) \neq \mathcal{D}es(o_1)$; o_1 and o_2 each imply P; o_3 implies $\neg P$; and $\mathcal{D}es(o_1) = x$, $\mathcal{D}es(o_2) = \mathcal{D}es(o_3) = y$. We require that $(o_1, o_2) =^{\mathrm{d}} (o_1, o_3)$, for obviously $x - y = x - y$. However, the justification for Definition 7.2 fails under these conditions:

$$(o_1, o_2) =^{\mathrm{d}} (o_1, o_3) \text{ iff } (o_1, P; o_3) \sim (o_2, P; o_1)$$

This holds just in case:

$$\mathcal{D}es(o_1 \,\&\, P).\mathcal{B}el(P) + \mathcal{D}es(o_3 \,\&\, \neg P).(1 - \mathcal{B}el(P)) = \mathcal{D}es(o_2 \,\&\, P).\mathcal{B}el(P) + \mathcal{D}es(o_1 \,\&\, \neg P).(1 - \mathcal{B}el(P))$$

Supposing $\mathcal{B}el(P) = ½$, this reduces to

$$½x + ½y = ½y + ½\mathcal{D}es(\bot)$$

It follows that $\mathcal{D}es(\bot) = x = \mathcal{D}es(o_1)$, which contradicts our initial assumption.

The only consistent way that Ramsey could have included impossible gambles in his system would have been to treat different impossible propositions as different objects of desire. However, the move from worlds to near-worlds in his characterisation of the outcome set $\mathcal{O}$ strongly suggests that he desired to avoid impossible gambles. And rightly so: restricting our attention to possible gambles seems the most plausible option. It is not obvious how we ought to treat preferences with respect to impossible gambles. For instance, it's implicit in Ramsey's system that if $o_1 \succcurlyeq o_2$, then $o_1 \succcurlyeq (o_1, P; o_2) \succcurlyeq o_2$. Without this assumption he is unable to show that $\mathcal{B}el$ is a credence function (see the proof of Theorem 8.3 in Appendix A). However, suppose that P is contingent, but we know that o_1 implies $\neg P$ and so $(o_1 \& P)$ cannot possibly obtain. In this case, it seems at least as plausible that $o_2 \succ (o_1, P; o_2)$ inasmuch as o_2 constitutes a desirable outcome—after all, we know we are *not* going to receive o_1 in the event that P and choosing $(o_1, P; o_2)$ only has leaves one with a chance $\mathcal{B}el(\neg P)$ of receiving o_2, so it would seem preferable to have o_2 for certain.

Thus, it looks as though Ramsey was implicitly assuming something even stronger than **RAM1**:

RAM1* For *every* $o \in \mathcal{O}$, there is at least one ethically neutral proposition P of credence ½ such that o is compatible P and $\neg P$

As I will argue shortly, **RAM1** is already too strong of an assumption for characterisational representationism to deal with. **RAM1*** is stronger still, and by a wide margin. Even where the former might be defended, the latter seems indefensible.

7.2.2 Problems with ethical neutrality

In looking at whether the notion of ethical neutrality is viable, we ought first to start with Ramsey's own definition:

Definition 7.6: Ethical neutrality (Ramsey's original)
P is ethically neutral iff (i) if P is atomic, then $w^P \sim w^{\neg P}$, for all pairs of worlds w^P, $w^{\neg P}$ identical with respect to all their atomic propositions except for P, (ii) if P is non-atomic, then all of Ps atomic truth arguments are ethically neutral

So, an atomic proposition P is *ethically neutral* for an agent iff any two possible worlds differing in their atomic propositions only in the truth of P are always equally valued by that agent, and ethical neutrality for non-atomic propositions is understood in terms of atomic propositions. Ramsey here demonstrates commitment to another aspect of Wittgensteinian atomism: every non-atomic proposition can be constructed from atomic propositions using truth-functional connectives. We are able to locate such a proposition, if it exists, by considering the agent's preferences over worlds. As just noted, for some gambles $(o_1, P; o_2)$ and $(o_2, P; o_1)$, Ramsey requires that o_1 and o_2 are compatible with both P and $\neg P$. If we suppose for simplicity that P is atomic, then o_1 and o_2 are near-worlds with respect to P. It follows from Ramsey's definition then that $(o_1 \,\&\, P) \sim (o_1 \,\&\, \neg P)$ and $(o_2 \,\&\, P) \sim (o_2 \,\&\, \neg P)$. It does *not* yet follow that $(o_1 \,\&\, P) \sim (o_1) \sim (o_1 \,\&\, \neg P)$, which Ramsey also required. However, we can take this as an unstated background assumption: if $(o_1 \,\&\, P) \sim (o_1 \,\&\, \neg P)$, then $(o_1 \,\&\, P) \sim (o_1) \sim (o_1 \,\&\, \neg P)$.

Sobel (1998, 241) has argued that there are few or no ethically neutral propositions in this sense. Consider the proposition *there are an even number of hairs on Dan Quayle*'*s head*. Sobel argues that this can be ethically neutral for 'almost no one':

> Though it is true that I do not care about Quayle's hair, there are worlds that differ regarding the truth of that proposition that, just because of that difference, differ in their values for me. I am thinking of worlds in which I have bet money on this proposition! The argument … can be readdressed to atomic propositions, if such there be, to the conclusion that *no* atomic proposition is Ramsey-ethically-neutral for any of us. (1998, 248)

There seem to be two concerns here. The first appears to be something like the following: for any proposition whatsoever, we should be able to find a set of otherwise similar possible worlds where we have entered into a bet conditional on that proposition with desirable outcomes if things turn out one way, and undesirable outcomes if things turn out another way. Since we care about the outcomes of the bet, we will value the relevant worlds differently. However, this objection seems to have no hold given Ramsey's view: the relevant worlds are supposed to differ at the atomic level *only* with respect to the proposition in question. In all other respects—including, importantly, the payouts for any bets we may enter into—the worlds are supposed to be identical.

The second and more obvious worry is that Ramsey's conception of ethical neutrality requires the assumption of logical atomism for its cogency. Ramsey built his theory upon the assumption of logical atomism so that he could make sense of the idea of two worlds differing *only* with respect to a particular proposition. The notion is of little use to contemporary philosophers who by and large reject that aspect of Wittgenstein's view. If we are to give $\succcurlyeq$ a plausible interpretation *qua* preference relation, we had better not build our account of that relation's objects on a now-defunct account of propositions.

In his atomism-free reconstruction of Ramsey's theorem, Bradley (2001) supplies the following definition, intended to achieve the same purpose:[88]

> **Definition 7.7: Ethical neutrality (atom-free)**
> P is ethically neutral iff for all propositions Q (that are compatible with both P and $\neg P$), $(P \& Q) \sim Q \sim (\neg P \& Q)$

Tautological and impossible propositions will be trivially ethically neutral according to this definition. Clearly, however, we are interested only in non-trivially ethically neutral propositions. A common suggestion is that propositions such as *the tossed coin will land heads* constitute ethically neutral propositions of credence ½. Part of the reason why we use coin tosses occasionally when making decisions is because we have no intrinsic interest in whether the coin lands heads or tails. If Q is something like *there are dogs*, and P is *the tossed coin will land heads*, then it seems plausible that $(P \& Q) \sim Q \sim (\neg P \& Q)$.

However, there are strong reasons to think that *no* contingent propositions will be ethically neutral in the sense of Definition 7.7, for any minimally rational subject. Let P be *the tossed coin will land heads*, and take Q to be the proposition (*the tossed coin will land heads* & *I receive $100000*) *or* (*the tossed coin will not land heads* & *I get kicked in the shins*). Q is obviously compatible with both P and $\neg P$. However, $(P \& Q)$ is equivalent to *the tossed coin will land heads* & *I receive $100000* while $(\neg P \& Q)$ is equivalent to *tossed coin will not land heads* & *I get kicked in the shins*. But for some very strange preference orderings, it's certainly not the case that $(P \& Q) \sim Q \sim (\neg P \& Q)$.

The point here generalises easily; there are no non-trivially ethically neutral propositions in this sense. Note that the issue here is not that no contingent proposition satisfies the definition *exactly*, while there may nevertheless be some propositions which *approximate* ethical neutrality. Rather, the upshot is that no proposition even comes *close* to satisfying the requirements of ethical neutrality. We will always be able to find countless many propositions Q that falsify the indifference requirements.

A refinement of Definition 7.7 might be useful. Instead of requiring $(P \& Q) \sim Q \sim (\neg P \& Q)$ for *all* Q compatible with both P and $\neg P$, Ramsey only requires the following:

[88] Definition 7.7 is a slight improvement upon the definition that Bradley actually gives in his paper, which does not include the restriction to propositions compatible with both P and $\neg P$. Without this restriction, any ethically neutral proposition, if it exists, has the same value as every necessary and every impossible proposition, and furthermore, the same value as every proposition Q which entails either P or $\neg P$. If we assume that the utility of a given proposition is determined by the (credence-weighted) utilities of its disjoint parts, then it will turn out on this definition that an ethically neutral proposition only exists if *all* propositions have precisely the same utility.

Definition 7.8: Ethical neutrality (atom-free, refined)
P is ethically neutral iff $o \sim (o \,\&\, P) \sim (o \,\&\, \neg P)$, for any outcome $o \in \mathcal{O}$ that is compatible with both P and $\neg P$

If there are no outcomes compatible with both P and $\neg P$, then P is trivially ethically neutral by this definition. Again, we can set such propositions aside; we are interested in non-trivially ethically neutral propositions. Definition 7.8 is weaker than Definition 7.7 because if Q is not in the outcome set $\mathcal{O}$, then there are no relevant gambles with Q as an outcome and we do not need to concern ourselves over whether $(P \,\&\, Q) \sim Q \sim (\neg P \,\&\, Q)$. More generally, if we assume that there are far fewer propositions in $\mathcal{O}$ than in $\mathcal{P}$, then the foregoing objection to Definition 7.7 is blocked. This will certainly be true if the outcomes in $\mathcal{O}$ are highly specific, as is the case in Ramsey's system.

With that said, it's still not obvious that any non-trivially ethically neutral propositions exist even in this weaker sense. Why should we suppose that there are *any* propositions P such that (non-trivially), $o \sim (o \,\&\, P) \sim (o \,\&\, \neg P)$ for all $o \in \mathcal{O}$ compatible with P and $\neg P$? And moreover, if **RAM1*** is being assumed, why should we suppose that for *every* $o \in \mathcal{O}$, we will find such propositions? Without knowing the exact nature of the outcome space $\mathcal{O}$, we cannot even know whether there *are* any outcomes compatible with both P and $\neg P$, for an arbitrarily chosen proposition P. Ramsey explicitly stipulates that there must be at least one pair of outcomes compatible with *some* ethically neutral proposition of credence ½ and its negation—but this stipulation is meaningless inasmuch as we do not already know what proposition that may be. Unfortunately, Ramsey's discussion leaves the nature of $\mathcal{O}$ quite vague, making the matter impossible to judge.

We can circumvent this concern by stipulating that $\mathcal{O}$ contains, for each of a very wide range of propositions in $\mathcal{P}$, outcomes that are undefined with respect to that proposition. But even then, Ramsey gives us little reason to suppose that ethically neutral propositions exist relative to a given agent's preference ordering—still less that there are any such propositions that satisfy Definition 7.1. **RAM1** clearly cannot be defended as a condition of rationality, and it does not follow from Ramsey's background assumption of the descriptive adequacy of CEU. Ramsey's aim in the first instance was to develop a procedure for the measurement of credences, so unlike other intended uses for decision-theoretic representation theorems he did not require his conditions to be constraints of practical rationality; nevertheless, if his process is to be viable then it ought at least be *applicable*. It may not be impossible for a rational agent to satisfy the condition, but we still require good reasons to believe that most do—yet no reasons are forthcoming.

A related issue regards Ramsey's proto-functionalist attempt to define credences in terms of his measurement procedure: a definition of credences which relies centrally on a dubitable and unjustified existential assumption is of very limited interest for characterisational representationism. Are we to suppose that agents who falsify **RAM1** do not have credences? Ultimately, given his reliance upon ethically neutral propositions, Ramsey's system was not sufficient to establish the main upshot of ‘Truth and Probability’: that the laws of probability provide for us the logic of partial belief. Even if it is understood in terms of Definition 7.8, **RAM1** is a very shaky foundation for a measurement procedure, and still worse for a characterisation of credences. Ramsey’s system fails to satisfy desideratum (**1**): it’s not plausible that his preference conditions are satisfied by many agents at all, if any.

Many expected utility representation theorems developed since ‘Truth and Probability’ have also made use of ethically neutral propositions, whether explicitly or implicitly. Davidson and Suppes (1956, see also Davidson, Suppes *et al.* 1957) develop a representation theorem similar to Ramsey’s wherein they explicitly characterise and assume the existence of ethically neutral propositions. Others make implicit appeal to ethically neutral propositions, in the sense that they figure in the intended interpretation of the formal system, rather than being formalised directly. In this capacity, for instance, we find ethical neutrality in the theorem of Debreu (1959), where $\succcurlyeq$ is defined on pairs of outcomes, which are understood as representing two-outcome gambles conditional on some ethically neutral P for which the agent has a credence of ½. Fishburne (1967) makes implicit appeal to ethically neutral propositions of credence ½ along very similar lines. Each of these works appear to require an understanding of ethical neutrality in something like the senses of Definition 7.7 or Definition 7.8 (each for essentially the same reason that Ramsey required the notion), and thus they inherit the problems associated with **RAM1**.

CHAPTER EIGHT

Ramsey without Ethical Neutrality

In this chapter, I will develop a representation theorem which comes close—both mathematically and conceptually—to Ramsey's original proposal for defining $\mathcal{B}el$ and $\mathcal{D}es$, but which does not require the appeal to ethically neutral propositions in any problematic sense. As we will see, the theorem to be developed also has several unique characteristics which make it particularly well-suited for the representation of ordinary agents.

Ramsey's proposal was to first construct a utility function $\mathcal{D}es$ using preferences over outcomes and gambles, following which the credence function $\mathcal{B}el$ could be defined over a set of propositions. We will follow a similar tact here. §8.1 outlines the key ideas behind the theorem. §8.2 then supplies the core theorem needed for the construction of $\mathcal{D}es$, while §8.3 provides the ensuing definition of $\mathcal{B}el$. §8.4 discusses the interpretation of a key part of the theorem, and §8.5 places it in connection with characterisational representationism.

8.1 Preliminaries

Recall that Ramsey's motivation for introducing the idea of ethical neutrality arises ultimately from his strategy for defining propositions of credence ½ and $=^{\mathrm{d}}$ (see §7.2.1). By adopting Ramsey's definitions, one is essentially forced to appeal to ethically neutral propositions or else fall into the trap of applying Naïve Expected Utility Theory to circumstances where it's both descriptively and normatively implausible. However, we are not forced to use Ramsey's definitions. It is possible to avoid introducing ethical neutrality in any of the problematic senses specified in §7.2.2, if we can develop alternative means of characterising propositions of credence ½ and $=^{\mathrm{d}}$.

8.1.1 Interpretations

Before we move on, it is worth saying a few words about the interpretation of the basic formal notions involved in the statement of the theorem: $\mathcal{O}$, $\mathcal{P}$, $\mathcal{G}$, $\succcurlyeq$, and a special relation $\rightharpoonup$. First of all, and unlike Ramsey, I will not assume that $\mathcal{O}$ should be comprised of either *worlds* or *near-worlds*. Instead, we will let $\mathcal{O}$ be an arbitrary set of propositions. In the formal treatment, $\mathcal{O}$ is essentially a set of points to be ordered by $\succcurlyeq$, and no special assumptions need to be made about its internal structure. Thus, the framework to be developed here is compatible with a wide range of theories about the nature of propositions (or the nature of *objects of thought* more generally), and we do not need to assume that logically equivalent propositions are identical elements in $\mathcal{O}$.

An arbitrary set of propositions $\mathcal{P}$ forms the domain of $\mathcal{B}el$. It would be possible to suppose that every proposition in $\mathcal{O}$ is in $\mathcal{P}$ (and *vice versa*), but this is not required for the theorem that follows and so will not be assumed. Importantly, *none* of the propositions in $\mathcal{P}$ need be very specific—in fact, they may be as fine-grained or coarse-grained as we like. Like $\mathcal{O}$, the formal treatment of $\mathcal{P}$ is compatible with many views on the nature of propositions, so logically equivalent propositions may form distinct elements of $\mathcal{P}$. For the purposes of constructing $\mathcal{B}el$, we will assume that $\mathcal{P}$ is closed under negation. This is a simplifying assumption only; see §8.3.3.

It is important for the result that follows that the propositions in $\mathcal{O}$ might stand in *implication* relations to the propositions in $\mathcal{P}$. I will make use of a special binary relation between propositions, denoted $\rightharpoonup$; and in the event that $P \rightharpoonup Q$ and $Q \rightharpoonup P$, we will write $P \rightleftharpoons Q$. As I will discuss in some detail in §8.4, I intend $P \rightharpoonup Q$ to mean that P *obviously implies* Q, where this is a non-transitive relation between pairs of propositions. For now, it should be assumed that for all P, Q,

(i) $\rightharpoonup$ is reflexive
(ii) $\rightharpoonup$ is neither symmetric nor antisymmetric
(iii) If $P \rightharpoonup Q$, then $P \rightleftharpoons (P \,\&\, Q)$
(iv) If $P \rightharpoonup Q$, then $P \vdash Q$

There are two important corollaries of (iv) to note: if $P \rightleftharpoons Q$, then $P \leftrightarrow Q$; and if $P \nvdash Q$, then $\neg(P \rightharpoonup Q)$.

The space of gambles $\mathcal{G}$ will be characterised as a proper subset of $\mathcal{O} \times \mathcal{P} \times \mathcal{O}$. The exact manner in which $\mathcal{G}$ is formalised is not especially important; however, it will be important that $\mathcal{G}$ is restricted to gambles $(o_1, P; o_2)$ such that:

(i) o_1 implies P and o_2 implies $\neg P$
(ii) If P is consistent, then o_1 is consistent, and if $\neg P$ is consistent, then o_2 is consistent
(iii) At least one of the pairs P and o_1 or $\neg P$ and o_2 must be non-equivalent

(i) helps to rules out the presence of what were earlier referred to as *impossible* gambles (Definition 7.5); a gamble will be found in $\mathcal{G}$ only if its outcomes imply the conditions in which they are supposed to obtain. (ii) then completes the removal of impossible gambles from $\mathcal{G}$, by ensuring (in combination with the first restriction) that if P and $\neg P$ are consistent then $(o_1 \,\&\, P)$ and $(\mathrm{o}_2 \,\&\, \neg P)$ are consistent. The conjunction of (i) and (ii) thus rules out the problematic state of affairs, discussed at the end of §7.2.1, where P is known to be possible but $(o_1 \,\&\, P)$ is known to be impossible, leading to $(o_1, P; o_2)$ being valued other than would be expected under the simple Naïve Expected Utility formula. (Note, though, that (ii) does not rule out gambles conditional on impossible propositions, nor does it rule out gambles with impossible outcomes—possible gambles may have impossible parts!)[89] Finally, (iii) rules out *trivial* gambles of the form $(P, P; \neg P)$, which will be discussed further in §8.2.2. I precisify these restrictions in **GRS1** below, and motivate them further in §8.4.

Our preference relation $\succcurlyeq$ will be defined on a space of outcomes $\mathcal{O}$ and a space of gambles $\mathcal{G}$ simultaneously. Since $\succcurlyeq$'s domain is $\mathcal{O} \cup \mathcal{G}$, the elements of $\mathcal{G}$ should be understood in a manner commensurate with those found in $\mathcal{O}$, lest $\succcurlyeq$ is given a highly disjunctive interpretation. Since $\mathcal{O}$ is an arbitrary collection of propositions, $\succcurlyeq$ on $\mathcal{O}$ is best thought of as a mentalistic preference relation, and $\succcurlyeq$ on $\mathcal{G}$ should be treated similarly. In particular, I would suggest that $(o_1, P; o_2) \succcurlyeq (o_3, Q; o_4)$ holds relative to a subject S just in case S would prefer (under considered reflection) the truth of *that she has accepted a gamble that returns o_1 if P, o_2 otherwise* to the truth of *that she has made a gamble that returns o_3 if Q, o_4 otherwise*.[90]

As noted in §7.1.1, such preferences would reliably correspond to a disposition to *choose* a gamble $(o_1, P; o_2)$ over $(o_3, Q; o_4)$ inasmuch as the subject accurately represents the gambles on offer—but we should not presume that she always does. $\succcurlyeq$ on $\mathcal{O} \cup \mathcal{G}$ cannot be given a behavioural reading independent of substantive (and implausible) assumptions about agents' doxastic states. The theorem to be developed, therefore, will not satisfy the naturalistic desideratum (5).

8.1.2 The basic strategy

The key idea of the two theorems to be developed is that, while Ramsey used the same *outcomes* in two distinct gambles $(o_1, P; o_2)$ and $(o_2, P; o_1)$ to define what it is for a proposition to have credence ½, his doing so was unnecessary: it's enough if we instead use outcomes with exactly the same *desirability*. That is, suppose that $o_1 \sim o'_1$ and $o_2 \sim$

[89] I am assuming that $(o_1, P; o_2)$ corresponds to a pair of *subjunctive* conditionals, $(P \mathrel{\Box\!\!\rightarrow} o_1) \,\&\, (\neg P \mathrel{\Box\!\!\rightarrow} o_2)$. I also assume that counterfactuals with impossible antecedents are vacuously true, as they are on a standard semantics for counterfactuals. Given this, every gamble in $\mathcal{G}$ corresponds to a *possible* conjunction of counterfactuals (while every impossible gamble corresponds to an *impossible* conjunction). I do not place very much weight on either of these assumptions—if some subjunctive conditionals with impossible antecedents are false, or if indicative conditionals are preferred and these admit of a quite distinct semantics, then further conditions can be placed on $\mathcal{G}$ to fix on the appropriate set.

[90] Note that it would be possible to interpret each element o in $\mathcal{O}$ as a gamble for o conditional on an obvious logical truth $\top$; i.e., as $(o, \top; x)$, where x is any arbitrary proposition. It would not seem implausible to suppose that $(o, \top; x) \succcurlyeq (o', \top; x)$ iff $o \succcurlyeq o'$.

o'_2, $\neg(o_1 \sim o_2)$, o_1 and o'_2 each imply P, while o_2 and o'_1 each imply $\neg P$, and finally, $(o_1, P; o_2) \sim (o'_2, P; o'_1)$. For now, I will continue to assume (as Ramsey did) that if $o \vdash P$, then $o \sim (o \;\&\; P)$; I will weaken this assumption in §8.4. Given this, and given the Ramseyan background assumption that "we act in the way that we think most likely to realize the objects of our desires" (§7.1.1), this situation is possible only if:

$$\mathcal{D}es(o_1).\mathcal{B}el(P) + \mathcal{D}es(o_2).(1 - \mathcal{B}el(P)) = \mathcal{D}es(o'_2).\mathcal{B}el(P) + \mathcal{D}es(o'_1).(1 - \mathcal{B}el(P))$$

Since $o_1 \sim o'_1$ and $o_2 \sim o'_2$, we know $\mathcal{D}es(o_1) = \mathcal{D}es(o'_1) = x$ and $\mathcal{D}es(o_2) = \mathcal{D}es(o'_2) = y$; and because $\neg(o_1 \sim o_2)$, we know that $x \neq y$. Let $\mathcal{B}el(P) = z$. We are left with:

$$xz + y(1 - z) = yz + x(1 - z)$$

Regardless of the specific values of x and y, this is possible only if $z = (1 - z)$; thus, $\mathcal{B}el(P) = ½$. There is no reason to require that P is ethically neutral.

Making the foregoing modifications forces a number of further changes to the basic formal system Ramsey developed. There are two particularly important changes that I will note here, before laying out the main theorem in full. First, we can no longer employ Ramsey's definition of $=^{\mathrm{d}}$. (Instead of defining $=^{\mathrm{d}}$, I will instead define $\geq^{\mathrm{d}}$.) However, we can employ the same trick as was just noted to avoid any appeal to ethical neutrality: there is no reason why $(o_1, o_2) \geq^{\mathrm{d}} (o_3, o_4)$ must be defined using o_1, o_2, o_3, and o_4 *in particular*. It's enough if we use outcomes with exactly the same desirability. And for that matter, there is no reason why we need to use the same proposition in both gambles, so long as we use a proposition of credence ½. Instead, we can say $(o_1, o_2) \geq^{\mathrm{d}} (o_3, o_4)$ holds iff, for each $(o'_1, P; o'_4), (o'_2, P'; o'_3) \in \mathcal{G}$ where P and P' are both of credence ½,

$$(o'_1, P; o'_4) \succcurlyeq (o'_2, P'; o'_3)$$

The reasoning behind this is essentially identical to the reasoning behind Definition 7.2.

Secondly, we need to ensure that there are enough outcomes for the new definition of $\geq^{\mathrm{d}}$ to generally apply. That is, we need to assume that we will always be able to find the required gambles $(o'_1, P; o'_4)$ and $(o'_2, P'; o'_3)$ in $\mathcal{G}$. This is not obviously going to be the case, given the earlier noted restriction on $\mathcal{G}$. In effect, we need to assume that for every pair o_1 and o_2, there will always exist at least one proposition P of credence ½ such that for some $o'_1 \sim o_1$ and $o'_2 \sim \mathrm{o}_2$, o'_1 implies P and o'_2 implies $\neg P$. This assumption implies that every *value* (see Definition 7.3) contains multiple members, and that at least two of these members will disagree with respect to some proposition P of credence ½. In effect, this assumption replaces Ramsey's condition **RAM1**; it is formalised as **GRS2** below.

8.2 Generalising Ramsey's system

I now develop a representation theorem for the construction of an interval scale $\mathcal{D}es$ on $\mathcal{G}$ and $\mathcal{O}$ such that for all $x, y \in \mathcal{O} \cup \mathcal{G}$ and all $o_1, o_2, o_3, o_4 \in \mathcal{O}$,

$$x \succcurlyeq y \text{ iff } \mathcal{D}es(x) \geq \mathcal{D}es(y)$$
$$(o_1, o_2) \geq^{\mathrm{d}} (o_3, o_4) \text{ iff } \mathcal{D}es(o_1) - \mathcal{D}es(o_2) \geq \mathcal{D}es(o_3) - \mathcal{D}es(o_4)$$

I will begin with a statement of the definitions, preference conditions, and ensuing representation theorem (§8.2.1), after which follows a discussion of each of the preference conditions (§8.2.2).

8.2.1 Main representation theorem

In what follows, I have adopted the notational convention that sameness of subscript for outcomes implies sameness of desirability (but the reverse need not hold). For instance, it should be assumed in all that follows that o'_1 and o''_1 each refer to outcomes with the same desirability as o_1 (i.e. $o_1 \sim o'_1$ and $o'_1 \sim o''_1$). It should not be assumed, however, that either o'_1 or o''_1 is necessarily distinct from o_1. Likewise, $(o_1, P; o_2)$ should be understood as a variable for gambles with outcome o_1 if P, o_2 otherwise; and $(o'_1, P; o'_2)$ for gambles conditional on P with outcomes *equal in value* to o_1 and o_2. Again, the pair $(o_1, P; o_2)$ and $(o'_1, P'; o'_2)$ need not be distinct.

We first define the set of propositions of credence ½:

Definition 8.1: Π
$\mathbf{\Pi} = \{P \in \mathcal{P}$: there are $o_1, o_2 \in \mathcal{O}$ such that $(o_1, P; o_2), (o'_2, P; o'_1) \in \mathcal{G}$, $\neg(o_1 \sim o_2)$, and $(o_1, P; o_2) \sim (o'_2, P; o'_1)\}$

Henceforth, I will use π, π', and so on, to designate propositions within $\mathbf{\Pi}$. It shouldn't be assumed that $\pi \neq \pi'$. Given this, I will use $(o_1, \pi; o_2)$ specifically for gambles conditional on *some* π in $\mathbf{\Pi}$ (with outcomes o_1 and o_2).

We can now define $\geq^{\mathrm{d}}$:

Definition 8.2: $\geq^{\mathrm{d}}$
$(o_1, o_2) \geq^{\mathrm{d}} (o_3, o_4)$ iff $(o'_1, \pi; o'_4) \succcurlyeq (o'_2, \pi'; o'_3)$ for all $(o'_1, \pi; o'_4), (o'_2, \pi'; o'_3) \in \mathcal{G}$

For the purposes of characterizing the Archimedean condition, we will also need to define a *strictly bounded standard sequence*. We can break this notion down into two concepts:

Definition 8.3: Standard sequence

$o_1, o_2, \ldots, o_i, \ldots$ is a standard sequence iff (i) for all $(o'_2, \pi; o'_1), (o''_1, \pi'; o'''_1) \in \mathcal{G}$, $\neg((o'_2, \pi; o'_1) \sim (o''_1, \pi'; o'''_1))$, and (ii) for every o_i, o_{i+1} in the sequence, $(o'_{i+1}, \pi; o'_2) \sim (o'_1, \pi'; o'_i)$ for all $(o'_{i+1}, \pi; o'_2), (o'_1, \pi'; o'_i) \in \mathcal{G}$

In light of the preference conditions to be characterised shortly, it will turn out that $o_1, o_2, \ldots, o_i, \ldots$ is a standard sequence iff $(o_2, o_1) \neq^d (o_1, o_1)$ and $(o_{i+1}, o_i) =^d (o_2, o_1)$ for all o_i, o_{i+1} in the sequence. So, for instance, the sequence o_1, o_2, o_3, o_4 is a standard sequence just in case:

$$(o_2, o_1) \neq^d (o_1, o_1) \text{ and } (o_4, o_3) =^d (o_3, o_2) =^d (o_2, o_1)$$

The idea, of course, is that the (nonzero) difference in desirability between any two adjacent members in the sequence is always equal to the difference in desirability between any other two adjacent members.

Definition 8.4: Strictly bounded standard sequence

$o_1, o_2, \ldots, o_i, \ldots$ is a strictly bounded standard sequence iff $o_1, o_2, \ldots, o_i, \ldots$ is a standard sequence and there exists $o_a, o_b \in \mathcal{O}$ such that for all o_i in the sequence, $(o'_a, \pi; o'_i) \succ (o'_1, \pi'; o'_b)$ and $(o''_i, \pi''; o''_b) \succ (o''_a, \pi'''; o''_1)$, for all $(o'_a, \pi; o'_i), (o'_1, \pi'; o'_b), (o''_i, \pi''; o''_b), (o''_a, \pi'''; o''_1) \in \mathcal{G}$

In other words, any standard sequence $o_1, o_2, \ldots, o_i, \ldots$ is strictly bounded if there are o_a, $o_b \in \mathcal{O}$ such that for any o_i in the sequence, $(o_a, o_b) >^d (o_i, o_1) >^d (o_b, o_a)$. Essentially, regardless of the size of the interval between o_i and o_1, we can find outcomes in $\mathcal{O}$ that are spaced even further apart.

The coherence of the foregoing definitions will be ensured by the conditions **GRS1–9**, which we can now specify.[91]

Definition 8.5: Generalised Ramsey structures

$<\mathcal{O}, \mathcal{P}, \mathcal{G}, \succcurlyeq>$ is a *generalised Ramsey structure* iff $\mathcal{O}$ and $\mathcal{P}$ are non-empty sets of propositions, $\mathcal{G} \subseteq \mathcal{O} \times \mathcal{P} \times \mathcal{O}$, $\succcurlyeq$ is a binary relation on $\mathcal{O} \cup \mathcal{G}$, and for all $o_1, o_2 \in \mathcal{O}$, all sequences $o_1, o_2, \ldots, o_i, \ldots \in \mathcal{O}$, all $P \in \mathcal{P}$, and all $(o_1, P; o_2), (o_1, \pi; o_2), (o'_2, \pi'; o'_1), (o_1, \pi; o_4), (o_2, \pi'; o_3), (o_3, \pi''; o_6), (o_4, \pi'''; o_5) \in \mathcal{G}$, the following nine conditions hold:

[91] In (Elliott forthcoming), I define a 'generalised Ramsey structure' in a slight different manner than I have done here. The primary difference involves a change to the characterisation of $\mathcal{O}$ (which is taken here as a set of consistent propositions rather than a set of worlds), which necessitates slightly different versions of **GRS1** and **GRS2**.

GRS1 $(o_1, P; o_2) \in \mathcal{G}$ iff (i) $o_1, o_2 \in \mathcal{O}$, (ii) $P \in \mathcal{P}$, (iii) $o_1 \rightharpoonup P$ and $o_2 \rightharpoonup \neg P$, (iv) either $P \nvdash o_1$ or $\neg P \nvdash o_2$, and (v) if P is consistent, then o_1 is consistent, and if $\neg P$ is consistent, then o_2 is consistent

GRS2 For every pair $o_1, o_2 \in \mathcal{O}$, there exists a $\pi \in \mathbf{\Pi}$ such that for some $o'_1, o'_2 \in \mathcal{O}$, (i) $o'_1 \rightharpoonup \pi$ and $o'_2 \rightharpoonup \neg\pi$, (ii) either $\pi \nvdash o'_1$ or $\neg\pi \nvdash o'_2$, and (iii) if π is consistent, o_1 is consistent, and if $\neg\pi$ is consistent, o_2 is consistent

GRS3 $\succcurlyeq$ on $\mathcal{O} \cup \mathcal{G}$ is a weak ordering

GRS4 If $(o_1, \pi; o_2), (o'_2, \pi'; o'_1) \in \mathcal{G}$, then $(o_1, \pi; o_2) \sim (o'_2, \pi'; o'_1)$

GRS5 If $(o_1, \pi; o_4) \succcurlyeq (o_2, \pi'; o_3)$ and $(o_3, \pi''; o_6) \succcurlyeq (o_4, \pi'''; o_5)$, then, for all $(o'_1, \pi^*; o'_6)$, $(o'_2, \pi^+; o'_5) \in \mathcal{G}$, $(o'_1, \pi^*; o'_6) \succcurlyeq (o'_2, \pi^+; o'_5)$

GRS6 For every triple $o_1, o_2, o_3 \in \mathcal{O}$, there is a $o_4 \in \mathcal{O}$ such that for some $(o'_1, \pi; o'_3)$, $(o_4, \pi'; o'_2) \in \mathcal{G}$, $(o'_1, \pi; o'_3) \sim (o_4, \pi'; o'_2)$

GRS7 If $o_1, o_2, \ldots, o_i, \ldots$ is a strictly bounded standard sequence, it is finite

GRS8 $o_1 \succcurlyeq o_2$ iff for all $(o'_1, P; o'_2) \in \mathcal{G}$, $o_1 \succcurlyeq (o'_1, P; o'_2) \succcurlyeq o_2$

GRS9 For each $(o_1, P; o_2) \in \mathcal{G}$, there is a $o_3 \in \mathcal{O}$ such that $(o_1, P; o_2) \sim o_3$

We can now state the main representation theorem:

Theorem 8.1: Generalised Ramseyan utility

If $<\mathcal{O}, \mathcal{P}, \mathcal{G}, \succcurlyeq>$ is a generalised Ramsey structure then there is a function $\mathcal{D}es$: $\mathcal{O} \cup \mathcal{G} \mapsto \mathbb{R}$ such that for all $x, y \in \mathcal{O} \cup \mathcal{G}$ and all $o_1, o_2, o_3, o_4 \in \mathcal{O}$,

(i) $x \succcurlyeq y$ iff $\mathcal{D}es(x) \geq \mathcal{D}es(y)$

(ii) $(o_1, o_2) \geq^d (o_3, o_4)$ iff $\mathcal{D}es(o_1) - \mathcal{D}es(o_2) \geq \mathcal{D}es(o_3) - \mathcal{D}es(o_4)$

Furthermore, $\mathcal{D}es$ is unique up to positive linear transformation

A proof is provided in Appendix A. The strategy behind the proof is closely connected to Ramsey's process; *viz*., given the agent's preferences over outcomes and gambles, we first determine the relation $\geq^d$ between pairs of outcomes and on that basis construct an interval scale measurement of the agent's preferences. The most important step here is to establish that if $<\mathcal{O}, \mathcal{P}, \mathcal{G}, \succcurlyeq>$ is a generalised Ramsey structure, then $<\mathcal{O} \times \mathcal{O}, \geq^d>$ is an *algebraic difference structure*:

Definition 8.6: Algebraic difference structure

$<\mathcal{X} \times \mathcal{X}, \succcurlyeq^*>$ is an *algebraic difference structure* iff $\mathcal{X}$ is non-empty, $\succcurlyeq^*$ is a binary relation on $\mathcal{X} \times \mathcal{X}$, and for all $x_1, x_2, x_3, x_4, x'_1, x'_2, x'_3 \in \mathcal{X}$, and all sequences $x_1, x_2, \ldots, x_i, \ldots \in \mathcal{X}$, the following five conditions hold:

ADS1 $\succcurlyeq^*$ on $\mathcal{X} \times \mathcal{X}$ is a weak ordering

ADS2 If $(x_1, x_2) \succcurlyeq^* (x_3, x_4)$, then $(x_4, x_3) \succcurlyeq^* (x_2, x_1)$

ADS3 If $(x_1, x_2) \succcurlyeq^* (x_4, x_5)$ and $(x_2, x_3) \succcurlyeq^* (x_5, x_6)$, then $(x_1, x_3) \succcurlyeq^* (x_4, x_6)$

ADS4 If $(x_1, x_2) \succcurlyeq^* (x_3, x_4) \succcurlyeq^* (x_1, x_1)$, then there exist $x_5, x_6 \in \mathcal{X}$ such that $(x_1, x_5) \sim^* (x_3, x_4) \sim^* (x_6, x_2)$

ADS5 If $x_1, x_2, \ldots, x_i, \ldots$ is such that $(x_{i+1}, x_i) \sim^* (x_2, x_1)$ for every x_i, x_{i+1} in the sequence, $\neg((x_2, x_1) \sim^* (x_1, x_1))$, and there exist $x', x'' \in \mathcal{X}$ such that $(x', x'') \succ^* (x_i, x_1) \succ^* (x'', x')$ for all x_i in the sequence, then it is finite

This allows us to invoke the following theorem:

Theorem 8.2: Algebraic difference measurement
If $<\mathcal{X} \times \mathcal{X}, \succcurlyeq^*>$ is an algebraic difference structure, then there exists a real-valued function $\mathcal{F}$ on $\mathcal{X}$ such that, for all $x_1, x_2, x_3, x_4 \in \mathcal{X}$,

(i) $(x_1, x_2) \succcurlyeq^* (x_3, x_4)$ iff $\mathcal{F}(x_1) - \mathcal{F}(x_2) \geq \mathcal{F}(x_3) - \mathcal{F}(x_4)$

Furthermore, $\mathcal{F}$ is unique up to positive linear transformation

For a proof of Theorem 8.2, see (Krantz, Luce *et al.* 1971, Ch. 4).

8.2.2 Generalised Ramsey structures

We now turn to a discussion of the conditions **GRS1–9** before looking at how to derive the credence function $\mathcal{B}el$. Though none of the conditions are identical to any of Ramsey's, many of them bear a close resemblance to the conditions and assumptions mentioned in his paper. It is worth noting that none of the conditions are intended to be independently plausible *qua* norms of practical rationality, though at least a few may seem to have this status.[92] As with Ramsey's formal system, the goal here is to establish conditions for the possibility of utility measurement under the assumption of the broad descriptive adequacy of something like expected utility theory—we aren't directly interested in establishing foundations for a prescriptive decision theory.

The purely structural condition **GRS1** does not correspond to any of Ramsey's conditions or any of the further assumptions he mentions. It is worth saying a few words about the fact that **GRS1** requires that $(o_1, P; o_2)$ is in $\mathcal{G}$ only if $P \nVdash o_1$ or $\neg P \nVdash o_2$. This assumption is *not* needed to prove Theorem 8.1, but it's important nonetheless. In particular, stating **GRS1** as such will help us to avoid a conflict between **GRS8** and a plausible

[92] Plausibly, **GRS3**, **GRS4**, and **GRS8** are constraints of practical rationality. I am inclined to take **GRS5** as a rationality constraint, though this is difficult to justify without presupposing the norm of expected utility maximisation. The status of the Archimedean condition **GRS7** is unclear, though representation theorems that forego an Archimedean condition can be developed, e.g., (Bartha 2007). The existential conditions **GRS2**, **GRS6**, and **GRS9** are not plausibly rationality constraints, but there is also a sense in which they are less important *vis-à-vis* the T-representability of $\succcurlyeq$ on $\mathcal{O} \cup \mathcal{G}$—namely, to the extent that they fail, $\mathcal{B}el$ and $\mathcal{D}es$ may be undefined for some propositions but not necessarily all (or even most). **GRS1** is a purely structural condition, and places no constraints on any agent whether ideal or not.

intuition about preferences over trivial gambles. Suppose, in particular, that $o_1 = P$ and $o_2 = \neg P$, so given the proposed interpretation of gambles, (o_1, P; o_2) represents *that S has accepted a gamble that returns P if P, ¬P otherwise*. **GRS8** asserts that the value of (o_1, P; o_2) should be somewhere between the values of o_1 and o_2. But now suppose that we have another gamble, (o_3, Q; o_4), where $o_3 = Q$, and $o_4 = \neg Q$; and suppose also that $o_1 \succ o_2 \succ o_3 \succ o_4$. Both ($o_1$, P; o_2) and (o_3, Q; o_4) represent utterly uninteresting prospects, and it would seem only rational to be indifferent between the two. This is, however, ruled out by **GRS8**, which now requires that $(o_1, P; o_2) \succ (o_3, Q; o_4)$.

GRS1 removes these kinds of *trivial* gambles from consideration. And this seems to be as it should be—trivial gambles of the form (P, P; $\neg P$) aren't really gambles at all—in a choice between two trivial gambles absolutely nothing is risked (or gained) one way or another. There is, therefore, no reason to consider one's credences and utilities regarding P and $\neg P$: trivial gambles are a special case where credences over the gamble's conditions and utilities for the gambles outcomes are irrelevant. On the other hand, where either $P \nvDash o_1$ or $\neg P \nvDash o_2$, (o_1, P; o_2) represents an *interesting* choice, and one where decision-makers' credences in P/$\neg P$ and utilities for o_1/o_2 seem very relevant.

In light of **GRS1**, **GRS2** essentially asserts that for every pair of outcomes o_1 and o_2, we will find at least one gamble in $\mathcal{G}$ conditional on some π in $\mathbf{\Pi}$ with outcomes equal in value to o_1 and o_2. It plays a very similar foundational role to **RAM1**; it is involved in most of the major steps of the proof of Theorem 8.1. However, **GRS2** is by far the more plausible condition. It implies the existence of a set of propositions, $\mathbf{\Pi}$, such that the agent prefers as though she believes each member of the set to degree ½, but *none* of them have to be ethically neutral in any of the senses defined in §7.2.2. Furthermore, unlike **RAM1***, **GRS2** does not require that every outcome has to be compatible with both the truth and falsity of at least one proposition in $\mathbf{\Pi}$. Given this, and independently of whatever might be said regarding its intrinsic plausibility, the use of **GRS2** as the basis for a representation theorem constitutes a substantial advance over Ramsey's system.

However, despite being more plausible than **RAM1** (and moreover **RAM1***), **GRS2** is nevertheless likely to be somewhat contentious. It implies, for instance, that *every* value $\underline{\underline{o}}$ contains at least two outcomes o and o' that differ with respect to their compatibility with some π in $\mathbf{\Pi}$. It's plausible that for *many* values—perhaps even most—we will be able to find such a proposition. Consider, for instance, the following situation. Our subject has no intrinsic interest in the outcomes of coin tosses. Let o be an arbitrary consistent outcome; and let π be the proposition *the next fair coin to be tossed lands heads*. Then, suppose that o' is (o & π), while o'' is (o & $\neg\pi$).[93] Plausibly, $o \sim o' \sim o''$, while o' (obviously) implies π and o'' (obviously) implies $\neg\pi$, but neither π nor $\neg\pi$ imply either o' or o''.

[93] If necessary, we might also suppose that no bets are made on the relevant coin toss, nor does its outcome affect history in any important way of interest to the decision-maker.

Importantly, **GRS2** is compatible with the possibility that some outcomes (or even *most* outcomes) o^*, $(o^* \,\&\, \pi)$ and $(o^* \,\&\, \neg\pi)$ might be valued quite differently.

The case just given suggests that for most outcomes we should be able to find a proposition of credence ½ which satisfies the conditions of **GRS2**. The condition seems to be at least approximately satisfied in this sense—for any outcome o, we should be able to find another outcome which is equivalent in all respects that the agent cares about but for the event of a fair coin toss. But it's still not obviously the case that this holds for *every* value $\underline{o}$. Perhaps there are some outcomes which are unique in their desirability ranking, being equal in value to no other; or perhaps there are some values which contain multiple outcomes, but none of which disagree with respect to any proposition of credence ½. This circumstance would seem to be rare if it occurs at all, and if so it would not be a devastating problem—it would primarily mean that sometimes, $\geq^{d}$ on $\mathcal{O} \times \mathcal{O}$ is undefined. Some pairs of outcomes might be left out of the $\geq^{d}$ comparison, but the relation would nevertheless still be a well-defined order on the others. It would likely be possible (though not without added complexity) to prove a weaker representation result, which leaves utility values for certain outcomes (and correspondingly, credence values for certain propositions) unspecified or within constrained intervals.

GRS3 corresponds closely to **RAM3**, and as we have seen, it is a standard necessary condition in decision-theoretic representation theorems. Although it is a very simple (and descriptively very plausible) condition, the role of **GRS4** is complex. No condition like it is in Ramsey's system, though amongst other things it plays many of the same roles as **RAM2**. In a manner of speaking, it says that the rational agent treats in the same way all prospects with similarly valued outcomes conditional on any proposition of credence ½. It tells us that we can substitute one outcome o_1 for another o'_1 within a gamble, or one proposition of credence ½ for another, so long as the outcomes have the same desirabilities and the substitution results in a possible gamble. So, for example, if o_1 and o'_1 have the same desirability and both are compatible with the propositions π and π', then $(o_1, \pi; o_2) \sim (o'_1, \pi'; o_2)$. It also allows that we can change the order of outcomes, in the sense that if $(o_1, \pi; o_2)$ and $(o'_2, \pi; o'_1)$ are both possible gambles, then $(o_1, \pi; o_2) \sim (o'_2, \pi; o'_1)$. **GRS4** helps to ensure the coherence of the definitions of $\mathbf{\Pi}$, $\geq^{d}$, and of $\mathcal{Bel}$.

GRS5 is designed to play the same role as **RAM4**. In light of the other conditions, it effectively asserts the reasonable proposition that $\geq^{d}$ is transitive, which is crucial for establishing that $<\mathcal{O} \times \mathcal{O}, \geq^{d}>$ satisfies **ADS1** and **ADS3** of Definition 8.6. The existential requirement **GRS6** is essentially a restatement of **RAM5**. Its role is limited to establishing that $<\mathcal{O} \times \mathcal{O}, \geq^{d}>$ satisfies **ADS4**, and is thus (like **ADS4**) a non-necessary structural condition. **GRS7** is the Archimedean condition; appropriately translated, it simply asserts that $<\mathcal{O} \times \mathcal{O}, \geq^{d}>$ satisfies **ADS5**.

GRS1–7 are sufficient to establish that $<\mathcal{O} \times \mathcal{O}, \geq^{d}>$ is an algebraic difference structure, which entails the existence of a real-valued function $\mathcal{Des}$ on $\mathcal{O}$ with the aforementioned properties. **GRS8–9** are then used to ensure that $\mathcal{Des}$ T-represents $\succcurlyeq$ on $\mathcal{O} \cup \mathcal{G}$. These final two conditions also play central roles in the construction of a credence function $\mathcal{Bel}$.

GRS8 does not correspond to any of Ramsey's stated conditions or any of the assumptions he otherwise mentions, though he clearly presupposed something like it. It states that the utility of a (non-trivial, possible) gamble $(o_1, P; o_2)$ sits somewhere weakly between the utilities of o_1 and o_2, which seems highly reasonable. This ensures that:

$$\mathcal{Des}(o_1) \geq \mathcal{Des}(o_2) \text{ iff } o_1 \succcurlyeq o_2$$

It also helps to ensure that $\mathcal{Bel}$ will never supply us with credence values of less than 0 or greater than 1.

The sole formal role of the existential condition **GRS9** is to ensure that we can extend $\mathcal{Des}$ on $\mathcal{O}$ to $\mathcal{O} \cup \mathcal{G}$; it is perhaps identical to what Ramsey intended for his **RAM7**. It necessitates the existence, for each gamble, of an outcome that is directly comparable with that gamble. Given the non-triviality of $\succ$ on $\mathcal{O} \cup \mathcal{G}$ (ensured by **GRS2**) and that, if $o_1 \succ o_2$, then $o_1 \succ (o_1, \pi; o_2) \succ o_2$, **GRS9** forces the set of outcomes to be infinite. In this respect, it's similar to Ramsey's **RAM6**, though it plays a quite different role than what Ramsey had intended for his condition. This is also likely to be a contentious condition; though here it is noteworthy that the assumption is not necessary for the main representation result. Other means of extending $\mathcal{Des}$ to $\mathcal{O} \cup \mathcal{G}$ are also likely possible in lieu of **GRS9**. Indeed, **GRS8** is alone enough to ensure that $\mathcal{Des}((o_1, P; o_2))$ sits somewhere weakly between $\mathcal{Des}(o_1)$ and $\mathcal{Des}(o_2)$. The failure of **GRS9** implies that, potentially but not necessarily, $\mathcal{Bel}$ as it will shortly be characterised may be undefined for some $P \in \mathcal{P}$.

8.3 Constructing *Bel*

Let us suppose that $<\mathcal{O}, \mathcal{P}, \mathcal{G}, \succcurlyeq>$ satisfies **GRS1–9**; our goal then is to construct a credence function $\mathcal{Bel}$ on $\mathcal{P}$ using $\mathcal{Des}$ on $\mathcal{O} \cup \mathcal{G}$, which combines with $\mathcal{Des}$ to form an expected utility T-representation of $\succcurlyeq$ on $\mathcal{O} \cup \mathcal{G}$. I will begin with a statement of the new conditions needed and the ensuing representation theorem (§8.3.1), after which I will note some interesting properties of the representation (§8.3.2), before discussing the new conditions—and possible weakenings thereof—in §8.3.3.

8.3.1 Secondary representation theorem

Closely following Ramsey's suggestion (§7.1.5), we can define $\mathcal{Bel}$ as follows:

Definition 8.7: $\mathcal{B}el$

For all $P \in \mathcal{P}$, if $o_1, o_2 \in \mathcal{O}$ are such that $\neg(o_1 \sim o_2)$ and $(o_1, P; o_2) \in \mathcal{G}$, then $\mathcal{B}el(P) = (\mathcal{D}es((o_1, P; o_2)) - \mathcal{D}es(o_2))/(\mathcal{D}es(o_1) - \mathcal{D}es(o_2))$

As with Ramsey's Definition 7.4, $\mathcal{B}el$ so-defined is unique.

There are three further conditions to add before we complete our construction of $\mathcal{B}el$. First of all, to ensure that there are enough gambles for $\mathcal{B}el(P)$ to always be defined, we will need to add the following structural condition to the previous nine preference conditions:

GRS10 For all $P \in \mathcal{P}$, there's at least one pair $o_1, o_2 \in \mathcal{O}$ such that (i) $\neg(o_1 \sim o_2)$, (ii) $o_1 \rightharpoonup P$ and $o_2 \rightharpoonup \neg P$, (iii) either $P \nvDash o_1$ or $\neg P \nvDash o_2$, and (iv) if P is consistent, then o_1 is consistent, and if $\neg P$ is consistent, then o_2 is consistent

Should **GRS10** fail, $\mathcal{B}el$ will be undefined for any proposition such that outcomes satisfying the stated conditions cannot be found.

Secondly, to ensure that $\mathcal{B}el(P)$ is always equal to $(1 - \mathcal{B}el(\neg P))$, we will also assume that:

GRS11 For all $(o_1, P; o_2), (o_2, \neg P; o_1) \in \mathcal{G}$, $(o_1, P; o_2) \sim (o_2, \neg P; o_1)$

GRS11 leads to condition (iv) of Theorem 8.3, but plays no other role besides this. It seems a very weak condition; it essentially states that the order in which outcomes are presented in a gamble makes no difference to their value. Another way to motivate **GRS11** would be to say that, despite being separate objects in $\mathcal{G}$, $(o_1, P; o_2)$ and $(o_2, \neg P; o_1)$ are mere notational variants representing the very same object of preference. If this is the case, then **GRS11** will fall out as a consequence of the intended interpretation of $\mathcal{G}$ and $\succcurlyeq$.

Thirdly, we will also need to assume the following special condition, stated in terms of $\mathcal{D}es$ rather than in terms of preferences, to ensure the coherence of Definition 8.7:

Condition 1: $\mathcal{B}el$ coherence

For all $(o_1, P; o_2), (o_3, P; o_4) \in \mathcal{G}$ where $\neg(o_1 \sim o_2)$ and $\neg(o_3 \sim o_4)$, $(\mathcal{D}es((o_1, P; o_2)) - \mathcal{D}es(o_2))/(\mathcal{D}es(o_1) - \mathcal{D}es(o_2)) = (\mathcal{D}es((o_3, P; o_4)) - \mathcal{D}es(o_4))/(\mathcal{D}es(o_3) - \mathcal{D}es(o_4))$

Condition 1 is a formal restatement of one of the conditions that Ramsey briefly mentions are required to ensure the coherence of the $\mathcal{B}el$ function (see §7.1.5). What it says can be visualised as follows. Definition 8.7 tells us that $\mathcal{B}el(P)$ is, say, 0.75, if it is the case that $o_1 \succ o_2$ and the value of the gamble $(o_1, P; o_2)$ sits exactly three quarters of the way from the values of o_2 to o_1. Condition 1 then tells us that for all o_3, o_4 such that $o_3 \succ o_4$, if the gamble (o_3, P; o_4) exists then it also sits three quarters of the way between o_4 and o_3 in the agent's desirability scale (and if $o_4 \succ o_3$, then (o_3, P; o_4) is one quarter of the distance between o_4 and o_3). This directly implies that the value $\mathcal{B}el(P)$ does not depend on which outcomes and gambles we choose to consider.

I have chosen to state Condition 1 in terms of $\mathcal{D}es$ as there is no apparent straightforward means of stating it purely in terms of preferences. Since $\mathcal{D}es$ is constructed entirely from preferences, Condition 1 is equivalent to *some* (perhaps infinitary) condition on preferences. Importantly, though, Condition 1's content is more transparent when expressed in terms of $\mathcal{D}es$, which requires of course that $\mathcal{D}es$ has already been characterised. Davidson and Suppes' (1956) condition A10 achieves the same purpose as my Condition 1 without referring to the intended T-representation, but only through a complicated series of definitions that serve to obscure its content—which is ultimately very similar to what Condition 1 says.

I will show in a moment that there is a way in which the definition of $\mathcal{B}el$ can be altered so as to remove the need for Condition 1—but for now, we now have the resources with which to construct an expected utility T-representation of $\succcurlyeq$ on $\mathcal{O} \cup \mathcal{G}$:

Theorem 8.3: Generalised Ramseyan credence and utility
If <$\mathcal{O}$, $\mathcal{P}$, $\mathcal{G}$, $\succcurlyeq$> is a generalised Ramsey structure where $\mathcal{P}$ is closed under negation, and **GRS10–11** and Condition 1 hold, then there is a function $\mathcal{D}es$: $\mathcal{O} \cup \mathcal{G} \mapsto \mathbb{R}$ and a function $\mathcal{B}el$: $\mathcal{P} \mapsto [0, 1]$ that for all x, $y \in \mathcal{O} \cup \mathcal{G}$, all $o_1, o_2, o_3, o_4 \in \mathcal{O}$, all ($o_1$, P; o_2) $\in \mathcal{G}$, and all $P \in \mathcal{P}$,

(i) $x \succcurlyeq y$ iff $\mathcal{D}es(x) \geq \mathcal{D}es(y)$

(ii) $(o_1, o_2) \geq^d (o_3, o_4)$ iff $\mathcal{D}es(o_1) - \mathcal{D}es(o_2) \geq \mathcal{D}es(o_3) - \mathcal{D}es(o_4)$

(iii) $\mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es(o_1).\mathcal{B}el(P) + \mathcal{D}es(o_2).(1 - \mathcal{B}el(P))$

(iv) $\mathcal{B}el(P) = 1 - \mathcal{B}el(\neg P)$

Furthermore, $\mathcal{B}el$ is unique and $\mathcal{D}es$ is unique up to positive linear transformation

A proof can be found in Appendix A.

8.3.2 Properties of $\mathcal{B}el$

It is important to note that $\mathcal{B}el$ need not be a probability function (though it is not inconsistent with the conditions that it could be); thus Theorem 8.3 is an NCU theorem. It is, however, a credence function, in that it maps propositions to some value within [0, 1]. The main restriction on $\mathcal{B}el$ is that if $\pi \in \mathbf{\Pi}$, then $\mathcal{B}el(\pi) = ½$, and there must be at least two $\pi \in \mathbf{\Pi}$.[94] I am inclined to take $\mathcal{B}el$'s potential lack of structure as a feature, not a bug. Plausibly, ordinary agents don't have probabilistically coherent (i.e., *additive*, *monotonic*) degrees of belief, so any representation of credences which requires such coherence is flawed.

The reason for $\mathcal{B}el$'s permissiveness is that **GRS1–11** and Condition 1 jointly place very few restrictions on preferences for gambles conditional on propositions outside of $\mathbf{\Pi}$. For instance, suppose that neither P nor Q are in $\mathbf{\Pi}$, P implies but is not equivalent to Q, $\neg(o_1 \sim o_2)$, and the agent is to rank the two gambles $(o_1, P; o_2)$ and $(o_1, Q; o_2)$. Only **GRS3**, **GRS8–9**, and Condition 1 can have any impact on how these gambles are ranked, as the other conditions are either purely existential or refer only to gambles conditional on propositions of credence ½. **GRS9** only asserts the existence of some o_3 and o_4 such that $o_3 \sim (o_1, P; o_2)$ and $o_4 \sim (o_1, Q; o_2)$, while **GRS8** only asserts that both $(o_1, P; o_2)$ and $(o_1, Q; o_2)$ must be valued somewhere between o_1 and o_2. Finally, Condition 1 only restricts the relative rankings of gambles conditional on the *same* proposition. All of these conditions, along with **GRS3**, can clearly be satisfied even if $(o_1, P; o_2) \succ (o_1, Q; o_2)$. Assuming all the other conditions to be satisfied, it follows immediately that if $(o_1, P; o_2) \succ (o_1, Q; o_2)$, then $\mathcal{B}el(P) > \mathcal{B}el(Q)$. Hence, $\mathcal{B}el$ in this instance is not a probability function, nor even a capacity.

Theorem 8.3 is thus compatible with an extremely wide range of credence functions. Indeed, $\mathcal{B}el$ is capable of assigning values of greater than 0 to impossible propositions, and less than 1 to necessary propositions. In §8.4, I will suggest a further condition which ensures that $\mathcal{D}es(P) = \mathcal{D}es(Q)$ and $\mathcal{B}el(P) = \mathcal{B}el(Q)$ if $P \rightleftharpoons Q$; thus, we can reasonably expect that any obvious impossibilities are assigned a credence of 0, and any obvious logical necessities a credence of 1. With further preference conditions, it's possible to ensure that $\mathcal{B}el$ satisfies particular structural properties, such as a weakened form of *monotonicity*: if $P \rightharpoonup Q$, then $\mathcal{B}el(Q) \geq \mathcal{B}el(P)$. For details, see (Elliott forthcoming).

Also important to note is that *none* of the propositions assigned values by $\mathcal{B}el$ (or $\mathcal{D}es$) need be very specific—in fact, they can for the most part be as fine-grained or as coarse-grained as we like. Because the formal treatment of propositions in Theorem 8.3 places so few constraints on the internal structure of either $\mathcal{O}$ or $\mathcal{P}$, we need not suppose anything as strong as, say, Jeffrey's assumption that $\mathcal{P}$ (minus a set of null propositions) forms a bottomless algebra with ever-increasingly fine-grained contents (§6.2).

Furthermore, $\mathcal{B}el$ and $\mathcal{D}es$ need not have wholly disjoint, non-overlapping domains. Indeed, $\mathcal{B}el$ and $\mathcal{D}es$ *can* be defined on precisely the same domain, or at least very similar domains. The main structural restriction here is **GRS10**, which is required if $\mathcal{B}el$ is to be

[94] The reasoning of §7.1.2 essentially counts as a proof that for all $\pi \in \mathbf{\Pi}$, $\mathcal{B}el(\pi) = ½$. That there is at least one proposition in $\mathbf{\Pi}$ follows immediately from **GRS2** and Definition 8.1, and that its negation (or something logically equivalent) is also in $\mathbf{\Pi}$ then follows from **GRS11**/**GRS11'**.

defined for all propositions in $\mathcal{P}$. (The falsity of this condition is compatible with $\mathcal{B}el$ being defined for *almost* all of $\mathcal{P}$.) Allowing that $\mathcal{P} = \mathcal{O}$ is consistent with **GRS10**, but does not imply it: setting $\mathcal{P} = \mathcal{O}$ ensures that there will always be $o_1, o_2 \in \mathcal{O}$ such that $o_1 \rightharpoonup P$, $o_2 \rightharpoonup \neg P$, for each $P \in \mathcal{P}$ (*viz.*, P and $\neg P$ themselves). However, this is not yet enough to guarantee the other conditions—for example, that the relevant o_1 and o_2 will be such that $\neg(o_1 \sim o_2)$, which **GRS10** also requires. I see no reason to think that the further conditions would not also be satisfied were we to assume that $\mathcal{P} = \mathcal{O}$—although if they are not, at most we would only require a few more propositions in $\mathcal{O}$ than in $\mathcal{P}$ (or $\mathcal{B}el$ could be left undefined for some propositions). Theorem 8.3 is unusual in this respect amongst multiset theorems, where $\mathcal{B}el$ and $\mathcal{D}es$ are usually *required* to have different domains.

8.3.3 Ways of weakening

There are (at least) two ways in which the conditions used to establish Theorem 8.3 can be weakened, leading to slightly different results. First of all, although we have assumed that $\mathcal{P}$ must be closed under negation, this is a simplifying assumption made to ensure that for any $P \in \mathcal{P}$, $\mathcal{B}el(\neg P)$ is defined. By making some reasonable assumptions about the character of $\rightharpoonup$, we can remove the closure condition and prove a slightly different result. Specifically, suppose that:

(a) If $P \rightleftharpoons Q$, then $R \rightharpoonup P$ only if $R \rightharpoonup Q$
(b) $\neg P \rightleftharpoons Q$ iff $\neg Q \rightleftharpoons P$

Then, replace **GRS11** with:

GRS11’ For all $P \in \mathcal{P}$, there is a $Q \in \mathcal{P}$ such that $\neg P \rightleftharpoons Q$, and if $(o_1, P; o_2), (o_2, Q; o_1) \in \mathcal{G}$, then $(o_1, P; o_2) \sim (o_2, Q; o_1)$

Given this, we can replace property (iv) of Theorem 8.3 with:

(iv′) For all $P, Q \in \mathcal{P}$, if Q is such that $\neg P \rightleftharpoons Q$, then $\mathcal{B}el(P) = 1 - \mathcal{B}el(Q)$[95]

This would avoid any need for assuming that if P is in $\mathcal{P}$, then so is $\neg P$, $\neg\neg P$, $\neg\neg\neg P$, and so on. In particular, suppose that P and $\neg P$ are in $\mathcal{P}$. Clearly, with respect to P, there is a

[95] Proof: If $P \in \mathcal{P}$ then there will be some $(o_1, P; o_2) \in \mathcal{G}$ by **GRS1** and **GRS10**. Our suppositions (i) and (ii) about $\rightharpoonup$ plus **GRS1** then imply that if the relevant Q exists in $\mathcal{P}$, then $(o_2, Q; o_1)$ will be in $\mathcal{G}$ as well. The first part of **GRS11’** then implies that the relevant Q can be found in $\mathcal{P}$, and the second part implies that $\mathcal{B}el(P) = 1 - \mathcal{B}el(Q)$, for essentially the reasons given in the proof of property (iv) of Theorem 8.3 in the Appendix A.

$Q \in \mathcal{P}$ such that $\neg P \rightleftharpoons Q$; namely, $\neg P$. So we can say $\mathcal{B}el(P) = 1 - \mathcal{B}el(\neg P)$. However, instead of going on to say that $\mathcal{B}el(\neg P) = 1 - \mathcal{B}el(\neg\neg P)$—in which case $\mathcal{B}el(\neg\neg P)$ would need to be defined, which would lead to $\mathcal{B}el(\neg\neg\neg P)$ needing to be defined, and so on—we can suppose that P is such that $\neg[\neg P] \rightleftharpoons P$, so $\mathcal{B}el(\neg P) = 1 - \mathcal{B}el(P)$. Thus, instead of requiring infinitely iterated negations, we can simply suppose that all (or most) of the propositions P in $\mathcal{P}$ can be paired with another proposition Q in $\mathcal{P}$ which is $\rightleftharpoons$-equivalent to P's negation.

Secondly, it is clear that Condition 1 is quite strong. Its satisfaction could only be expected of an agent who is extraordinarily consistent with respect to her preferences over gambles—in effect, it requires that for *all* the relevant gambles conditional on P, the agent has preferences as though she were a flawless expected utility maximiser with an infinitely precise credence in P. (It is because of Condition 1 that $\mathcal{B}el$ is a credence function, which is only capable of assigning point-like credence values to propositions.) This is more than we can expect of any ordinary subject. As it turns out, however, we can do without Condition 1 with some tweaks to the definition of $\mathcal{B}el$:

Definition 8.8: $\mathcal{B}el$*

For all $P \in \mathcal{P}$, $\mathcal{B}el^*(P) = [\lambda_1, \lambda_2]$ if and only if $[\lambda_1, \lambda_2]$ is the smallest interval such that for any $(o_1, P; o_2) \in \mathcal{G}$ where $\neg(o_1 \sim o_2)$, $(\mathcal{D}es((o_1, P; o_2)) - \mathcal{D}es(o_2)) / (\mathcal{D}es(o_1) - \mathcal{D}es(o_2)) \in [\lambda_1, \lambda_2]$

Note that $\mathcal{B}el^*$ will be unique, in the sense that for any P there is only one *smallest* interval $[\lambda_1, \lambda_2]$ such that $(\mathcal{D}es((o_1, P; o_2)) - \mathcal{D}es(o_2)) / (\mathcal{D}es(o_1) - \mathcal{D}es(o_2)) \in [\lambda_1, \lambda_2]$ for any $(o_1, P; o_2)$ where $\neg(o_1 \sim o_2)$. This is for essentially the same reason that $\mathcal{B}el$ is unique.

Here is the intuitive idea behind $\mathcal{B}el^*$. Definition 8.7 essentially says that $\mathcal{B}el(P) = 1/n$ just in case the agent treats *all* gambles conditional on P as though she assigns a credence of $1/n$ to P, in the sense that the value of $\mathcal{D}es$ for all gambles $(o_1, P; o_2)$ with $o_1 \succ o_2$ sit $1/n$ of the way between o_1 and o_2. Definition 8.8, on the other hand, allows for some variability in the agent's preferences with respect to gambles conditional on P, and $\mathcal{B}el^*$ represents that variation by means of an interval. For example, suppose that $o_1 \succ o_2 \succ o_3 \succ o_4$, and that on the one hand the agent's value for $(o_1, P; o_2)$ sits ½ way between her values for o_1 and o_2, while on the other hand her value for $(o_3, P; o_4)$ is ¼ of the way between o_3 and o_4. For simplicity, suppose first of all that $(o_1, P; o_2)$ and $(o_3, P; o_4)$ are the only gambles conditional on P. Then, $\mathcal{B}el^*(P)$ would be [¼, ½]. If there were one more gamble to consider—say, $(o_1, P; o_4)$—and its value sat ⅓ of the way between the values of its outcomes, then $\mathcal{B}el^*(P)$ would remain unchanged; however, if it was ⅕ of the way, then $\mathcal{B}el^*(P)$ would equal [⅕, ½].

Notice that if Condition 1 *is* satisfied, then $\mathcal{B}el(P) = n$ just in case $\mathcal{B}el^*(P) = [n, n]$. Thus, $\mathcal{B}el^*$ can be seen as a generalisation of $\mathcal{B}el$, the latter reducing to the former in the special case that Condition 1 holds. Importantly, though, Definition 8.8 does not require Condition 1 (or any other special conditions) to be satisfied in order for $\mathcal{B}el^*$ to be defined for *any* proposition in $\mathcal{P}$. Given **GRS11**, property (iv) of Theorem 8.3 would be replaced with:

(iv″) $\mathcal{B}el^*(P) = [\lambda_1, \lambda_2]$ iff $\mathcal{B}el^*(\neg P) = [1 - \lambda_2, 1 - \lambda_1]$

Note that dropping Condition 1 and adopting Definition 8.8 would mean that the expected utility T-representation of $\mathcal{D}es$ on $\mathcal{G}$ would also need to be altered slightly; in particular, instead of condition (iii) as stated in Theorem 8.3, we would now need to say that for each $(o_1, P; o_2) \in \mathcal{G}$,

(iii′) $\mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es(o_1).\lambda + \mathcal{D}es(o_2).(1 - \lambda)$, for some $\lambda \in \mathcal{B}el^*(P)$

The most plausible way to understand $\mathcal{B}el^*$ is to take it as providing us with a *limit* on any adequate measure of the agent's credences towards P, *given* her preferences and under the assumption that she at least approximately evaluates the utility of gambles according to their expected utility. In other words, I would suggest that $\mathcal{B}el^*(P) = [\lambda_1, \lambda_2]$ tells us that the agent's preferences constrain what her credence in P may be at least down to $[\lambda_1, \lambda_2]$, on the presupposition that she approximates the norm of expected utility maximisation. This reading of $\mathcal{B}el^*$ is compatible with a range of possibilities. For instance, $\mathcal{B}el^*(P) = [\lambda_1, \lambda_2]$ would be consistent with the agent having a sharp credence for P anywhere within $[\lambda_1 ,\lambda_2]$—in which case she is presumably somewhat inconsistent with respect to how she evaluates the utilities of gambles conditional on P. It is also compatible with the agent having *imprecise* credences accurately measured by some interval within $[\lambda_1, \lambda_2]$, including but not necessarily $[\lambda_1, \lambda_2]$ itself. I do not think either of these interpretations should be given priority over the others; at best, $\mathcal{B}el^*(P) = [\lambda_1, \lambda_2]$ should only be taken to mean that *whatever* the true measure of the agent's credences regarding P may be, it (most likely) sits somewhere within $[\lambda_1, \lambda_2]$. Further information would need to be considered to determine where *exactly* the agent's credences in P should be located.

A theorem without Condition 1 seems desirable for characterisational representationism, but I want to draw a more general lesson from the present discussion. Condition 1 is a strong requirement, but its strength is directly connected to the strict requirements that have been placed the intended T-representation of $\succcurlyeq$. Without the full strength of the condition, there would not exist *any* credence function $\mathcal{B}el$ such that for all $(o_1, P; o_2) \in \mathcal{G}$,

(iii) $\mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es(o_1).\mathcal{B}el(P) + \mathcal{D}es(o_2).(1 - \mathcal{B}el(P))$

In this sense, Condition 1 is *necessary* for the existence of the T-representation.

But even where these kinds of conditions fail, there will usually be a closely related result in the vicinity. Roughly put, there is in general some wriggle-room with respect to preference conditions—ways to loosen the rather strict requirements they impose—so long as we are prepared to live with a somewhat less precise T-representation.[96] $\mathcal{B}el^*$ is one instance of this possibility, but we should expect that there are more. In particular, it should be expected that the precision of $\mathcal{D}es$ is a consequence of the sometimes rather strict conditions imposed by **GRS1**–**9**, and it would be possible to loosen these conditions to arrive at a more general utility function for the T-representation of $\succcurlyeq$ on $\mathcal{O} \cup \mathcal{G}$.

8.4 Two interpretations of $\rightharpoonup$

I have allowed that logically equivalent propositions may be counted as distinct elements in $\mathcal{O}$ and $\mathcal{P}$, and for that reason $\mathcal{B}el$ and $\mathcal{D}es$ are capable of distinguishing between logical equivalencies. However, we must be very careful about what we say here, as much hangs on how we interpret $\rightharpoonup$.

Suppose that $o \rightharpoonup P$ means $o \vdash P$. As discussed in §7.2.1, if we make the Indifference to Equivalent Conjunctions assumption (that $P \sim (P \,\&\, Q)$ whenever $P \vdash Q$), then we can apply the Naïve Expected Utility formula to any gamble where the outcomes entail the conditions under which they obtain. This was an important (albeit implicit) background assumption behind Ramsey's Representation Conjecture—and as we will see—something similar is needed to underlie Theorem 8.3. Equating $\rightharpoonup$ with $\vdash$ comes with rather severe interpretational difficulties, however.

To get a grip on the central problem here, suppose first of all that $o_1 \vdash o_2$, and that o_1 and $(o_1 \,\&\, o_2)$ are in $\mathcal{O}$ and in $\mathcal{P}$. Given Indifference to Equivalent Conjunctions, it should be the case that $o_1 \sim (o_1 \,\&\, o_2)$. However, this is *not* implied by any of the conditions **GRS1**–**11**. Consistently with those conditions, then, the agent might prefer o_1 to $(o_1 \,\&\, o_2)$. But now suppose that $o_2 = P$ for some $P \in \mathcal{P}$. Theorem 8.3 then tells us that:

$$\mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es(o_1).\mathcal{B}el(P) + \mathcal{D}es(o_2).(1 - \mathcal{B}el(P))$$

But this is surely wrong. By Theorem 8.1, $\mathcal{D}es(o_1) \neq \mathcal{D}es(o_1 \,\&\, P)$, and because we wish to avoid applying Naïve Expected Utility Theory where that theory is inappropriate, the value of $(o_1, P; o_2)$ should be given by:

$$\mathcal{D}es(o_1 \,\&\, P).\mathcal{B}el(P) + \mathcal{D}es(o_2 \,\&\, \neg P).(1 - \mathcal{B}el(P))$$

[96] On this, see especially the discussion on coherent extendibility, §5.2.3.

Under this interpretation of $\rightharpoonup$, Theorem 8.3 would have us represent an agent who prefers o_1 to $(o_1 \& o_2)$ in a way which only seems appropriate given Indifference to Equivalent Conjunctions. **GRS1–11** are therefore consistent with a preference set which falsifies a central background assumption needed to motivate the theorem itself. Something has gone wrong.

One could introduce a further preference condition to avoid the foregoing worry. The weakest condition in the vicinity would be:

$$\text{If } o_1 \vdash o_2 \text{ and } o_1, (o_1 \& o_2) \in \mathcal{P}, \text{ and } o_2 = P \text{ for any } P \in \mathcal{P}, \text{ then } o_1 \sim (o_1 \& o_2)$$

This is not as strong as Indifference to Equivalent Conjunctions, but it would suffice to prevent the problematic state of affairs just discussed. However, it's *ad hoc* at best. Indeed, even requiring that $P \sim (P \& Q)$ whenever $P \vdash Q$ seems an odd restriction to impose—why not go all the way and assume that the agent does not distinguish between P and Q for the purposes of decision-making whenever P and Q are logically equivalent? After all, even the weaker condition imposes a kind of deductive infallibility upon the agent—an ability to always recognise when $o_1 \vdash o_2$ for arbitrary o_1 and o_2 satisfying the relevant conditions—and there seems to be no important difference between this kind of infallibility and the more general ability to determine the logical relationships between *any* pair of propositions that might be considered. So it seems that if $\rightharpoonup$ is taken to mean $\vdash$, Theorem 8.3 is only plausible for agents who always recognise and assign the same utilities and credences to logically equivalent propositions. We are left without a model for ordinary agents, who lack such deductive brilliance: in many cases, it might be *unobvious* when $o_1 \vdash o_2$.

Furthermore, the remarked upon flexibility of $\mathcal{Bel}$ becomes rather odd on this picture. If our subject is always able to recognise implication relations, then we might expect her credences to satisfy *at least* monotonicity—but we have seen that $\mathcal{Bel}$ need not be monotonic. It seems implausible to demand in the first place an extraordinary degree of rationality with respect to one domain ($\succcurlyeq$ and $\mathcal{Des}$), whilst at the same time representing that agent as highly irrational with respect to another closely related domain ($\mathcal{Bel}$). Inasmuch as we need to presuppose that the agent has some special kind of deductively infallibility to motivate Theorem 8.1, it had better not be the case that Theorem 8.3's $\mathcal{Bel}$ and $\mathcal{Des}$ functions represent the agent as being logically incompetent!

The cause of the problem is that Theorem 8.3 should only be used in cases where the Naïve Expected Utility formula is descriptively plausible. If $P \rightharpoonup Q$ is taken to mean $P \vdash Q$, then **GRS1** will imply that $\mathcal{G}$ includes gambles for which the Naïve Expected Utility formula is grossly inadequate for less-than-ideal agents. However, if we let $P \rightharpoonup Q$ mean that P *obviously implies* Q, then we might retain the plausibility of Theorem 8.3's representation without presupposing anything as strong as Indifference to Equivalent Conjunctions, while at the same time distinguish amongst some logically equivalent propositions. By ‘*P obviously implies Q*’, I mean that we can reasonably expect anyone capable of entertaining attitudes towards P and Q to *recognise* that P implies Q. For example, we can reasonably expect that anyone who understands P to know that P implies $(P \,\&\, P)$, but it may not be so obvious that P implies $(P \longrightarrow \neg((\neg P \vee Q) \,\&\, (P \,\&\, \neg Q)))$. I will have more to say on obvious implication shortly, but for now, let us see what can be done with this interpretation of $\rightharpoonup$.

We will continue to allow that logically equivalent propositions may form different elements in $\mathcal{P}$ and $\mathcal{O}$, but to avoid the earlier troubles we will assume that the following holds for all $P, Q \in \mathcal{P} \cup \mathcal{O}$:

> **Indifference between Obvious Equivalents**
> If $P \rightleftharpoons Q$, then (i) $P \sim Q$, and (ii) if $(o_1, P; o_2), (o'_1, Q; o'_2) \in \mathcal{G}$, then $(o_1, P; o_2) \sim (o'_1, Q; o'_2)$

To be clear, Indifference between Obvious Equivalents is *not* required to establish Theorem 8.3; in that sense, it is superfluous. Instead, it should be thought of as a restriction on the kinds of preference systems to which Theorem 8.3 can be reasonably supposed to apply.

And Indifference between Obvious Equivalents seems like an incredibly plausible assumption, both rationally and descriptively. Roughly, under the present interpretation of $\rightharpoonup$, Theorem 8.3 says that if P and Q are *obviously* equivalent, then (i) the agent in question is indifferent between P and Q, and (ii) she will also be indifferent between any two gambles of the form $(o_1, P; o_2)$ and $(o'_1, Q; o'_2)$, as each has a $\mathcal{B}el(P) = \mathcal{B}el(Q)$ likelihood of resulting in an outcome equal in value to o_1 and a $(1 - \mathcal{B}el(P)) = (1 - \mathcal{B}el(Q))$ chance of resulting in an outcome equal in value to o_2. Or, in more direct terms, *she does not distinguish between obviously equivalent propositions when forming her preferences over* $\mathcal{G} \cup \mathcal{O}$. We expect that this is how an ordinary agent would treat propositions that she *recognises* as being equivalent, so we can likewise expect that this is how she would treat propositions which are *obviously* equivalent (and hence recognised as such). Even if ordinary agents don’t live up to this very weak standard of rationality, it can hardly be doubted that they *approximate* the condition quite closely—and any agent who does not even come close to satisfying Indifference between Obvious Equivalents may perhaps be too irrational to have coherently measured credences and utilities in any case.

In the context of the other preference conditions, Indifference between Obvious Equivalents straightforwardly implies that $\mathcal{D}es(P) = \mathcal{D}es(Q)$ and $\mathcal{B}el(P) = \mathcal{B}el(Q)$ whenever $P \rightleftharpoons Q$. Propositions which are not obviously equivalent may, however, be assigned distinct values by $\mathcal{B}el$ and $\mathcal{D}es$. We have also assumed that if $P \rightharpoonup Q$, then $P \rightleftharpoons (P \,\&\, Q)$—so given Indifference between Obvious Equivalents,

If $P \rightharpoonup Q$, then $P \sim (P \,\&\, Q)$

This is substantially weaker than Indifference to Equivalent Conjunctions. More generally, we only need to suppose that the agent in question is deductively infallible with respect to *obvious* logical inferences—that is, with respect to the kinds of inferences that, by hypothesis, we can expect her to reliably make. Note also that Indifference between Obvious Equivalents only forces a kind of preference consistency between gambles conditional on obviously equivalent propositions whenever those gambles have the *same* or *equally valued* outcomes. Its satisfaction is therefore compatible with a failure to satisfy Condition 1, in which case Indifference between Obvious Equivalents implies that $\mathcal{B}el^*(P) = \mathcal{B}el^*(Q)$ whenever $P \rightleftharpoons Q$.

There is, then, the issue of specifying the *obvious implication* relation. There are clear cases in which P obviously entails Q; for instance, that *there are dogs* obviously implies that *there are things*; and *there are dogs and cats* obviously implies *there are cats*. And there are clear cases where P does not obviously entail Q; for instance, that *there are dogs* implies that *there are infinitely many primes*, but this is by no means obvious. And finally, there are also cases where an implication may be obvious to some, but not so obvious to others. It would be incongruous with characterisational representationism to presuppose knowledge of when S *recognises* that P implies Q—recognition is a kind of doxastic state that is far too close to what the characterisational representationist is aiming to explain—so it seems some notion of *objective obviousness* may be needed here.

In specifying $\rightharpoonup$, we may indeed have to impose some specification of *obviousness* from the outside, so to speak. There are some inferences which just *are* obvious, which most people recognise as obvious, and which *should* be obvious to anybody worthy of being called an agent, at least in *normal* conditions. The most plausible interpretation of $\rightharpoonup$ is that it represents *these* inferences. Here, we might appeal to a notion of minimal rationality as a constitutive norm of agency: part of what it is for S to be an agent at all *is* for S to be minimally rational, to respond appropriately to the evidence around her, and to make rational choices in light of that evidence.[97] Plausibly, to say that S is an agent is to presuppose that S at least comes close to satisfying some such criterion of rationality, at least under normal conditions. It seems only natural, then, that we would also suppose her to draw the obvious implications from the propositions she considers and to recognise obvious logical equivalences, *ceterus paribus*.

[97] Compare the principle of Charity, §4.2.

But we can say a little bit more about $\rightharpoonup$ than just this. In particular, Indifference between Obvious Equivalents can be reverse-engineered to provide a criterion of adequacy for any characterisation of $\rightharpoonup$. Roughly, the idea is that any specification of the *obvious implication* relation had better be such that all (or almost all) of the people in the relevant community (have preferences which suggest that they) recognise obvious logical equivalencies. It's reasonable to suppose that ordinary agents in normal circumstances don't distinguish between propositions that they *recognise* as being logically equivalent when forming their preferences. So, for any proposed specification of the *obvious* implication relation, it ought to be the case that:

> P *obviously implies* Q only if $P \vdash Q$, and for all (or most) members of the relevant community, $P \sim (P \mathbin{\&} Q)$, and if $(o_1, P; o_2)$, $(o'_1, (P \mathbin{\&} Q); o'_2) \in \mathcal{G}$, then $(o_1, P; o_2) \sim (o'_1, (P \mathbin{\&} Q); o'_2)$

Indeed, if P actually implies Q, and all (or almost all) of the people in the community don't seem to distinguish between P and $(P \mathbin{\&} Q)$ when forming their preferences, then in general the best explanation of this fact would be that P is obviously equivalent to $(P \mathbin{\&} Q)$, and so P obviously implies Q. The restriction to a 'relevant community' is intended to introduce some flexibility to the specification of $\rightharpoonup$ across different contexts and for different subjects. For example, what is obvious to mathematicians may not be obvious to the folk; and what is obvious to adults may not be obvious to children.

8.5 Limiting $\mathcal{G}$

It is worth noting, under the proposed interpretation of $\rightharpoonup$, $\mathcal{G}$ is in certain respects highly limited, having been reduced to just those non-trivial two-outcome gambles wherein the outcomes *obviously imply* the conditions under which they obtain. This does not seem to make the existential requirements (specifically, **GRS2** and **GRS10**) any more problematic than they would have been had $\rightharpoonup$ been interpreted as $\vdash$. However, the restriction on $\mathcal{G}$ does mean that information about preferences over *other* types of gambles is effectively ignored in the generation of $\mathcal{B}el$ and $\mathcal{D}es$—and there are *many* more possible gambles than can be found in $\mathcal{G}$. For one thing, there are gambles where the outcomes are merely *consistent* with their conditions, but don't *imply* them. The preference conditions of Theorem 8.3 are also consistent with all kinds of preference patterns over various *impossible* gambles.

Furthermore, there are also gambles to consider which have more than two possible outcomes. For instance, suppose that $\{P_1, P_2, P_3\}$ is a partition of logical space. Nothing about Theorem 8.3's preference conditions implies that the desirability of the slightly more complex gamble, $(o_1, P_1; o_2, P_2; o_3, P_3)$, must be a function of $\mathcal{B}el$ and $\mathcal{D}es$ as derived from the agent's preferences over two-outcome gambles. In particular, there is nothing to ensure that:

$$\begin{aligned}\mathcal{D}es((o_1, P_1; o_2, P_2; o_3, P_3)) \;=\; & \mathcal{D}es(o_1).\mathcal{B}el(P_1) + \mathcal{D}es(o_2).\mathcal{B}el(P_2) + \\ & \mathcal{D}es(o_3).\mathcal{B}el(P_3)\end{aligned}$$

It would be trivial to apply an *ad hoc* condition which ensures the above representation of $(o_1, P_1; o_2, P_2; o_3, P_3)$ and other finitely complex gambles; namely,[98]

> **Condition 2: Complex gamble consistency**
> For each $(o_1, P_1; \ldots; o_n, P_n)$, there is an outcome $o' \in \mathcal{O}$ such that $o' \sim (o_1, P_1; \ldots; o_n, P_n)$ and $\mathcal{D}es(o') = \mathcal{D}es(o_1).\mathcal{B}el(P_1) + \ldots + \mathcal{D}es(o_3).\mathcal{B}el(P_3)$

As with Condition 1, this is equivalent to *some* condition stated only in terms of $\succcurlyeq$, though likely an infinitely complex one. But, besides being *ad hoc*, the inclusion of Condition 2 would not alter the fact that $\mathcal{B}el$ and $\mathcal{D}es$ are derived *entirely* on the basis of preferences between two-outcome gambles. The real problem is not that we have no T-representation of $\succcurlyeq$ over a space of n-outcome gambles for $n > 2$, but that whatever representation we do have depends on preferences over such a restricted space.

This kind of limitation is problematic inasmuch as an agent's preferences over two-outcome gambles may not line up nicely with her preferences over more complex gambles to be. Suppose, for instance, that S's preferences for the gambles in $\mathcal{G}$ satisfy Indifference between Obvious Equivalents, **GRS1–11** and Condition 1, in which case it seems natural to interpret her as preferring between gambles according to a rule of expected utility maximisation with the $\mathcal{D}es$ and $\mathcal{B}el$ functions thus supplied by Theorem 8.3. However, suppose also that her preferences for *three*-outcome gambles, while wildly at odds with expected utility maximisation under $\mathcal{B}el$ and $\mathcal{D}es$, *would* be rationalised under $\mathcal{B}el^+$ and $\mathcal{D}es^+$. Which representation is the correct one, if any? Suppose further that her preferences between all three-or-more-outcome gambles would be rational under $\mathcal{B}el^+$ and $\mathcal{D}es^+$, and indeed her preferences for two-outcome gambles *not* in $\mathcal{G}$ would *also* be rational under $\mathcal{B}el^+$ and $\mathcal{D}es^+$. Surely, in this case, S's preferences over the gambles in $\mathcal{G}$ should be seen as an anomaly—whatever her credences and utilities may be, the vast majority of her preferences over the gambles *outside* of $\mathcal{G}$ would not make sense if we interpret her using $\mathcal{B}el$ and $\mathcal{D}es$.

I do not think that either of these issues should be taken to render Theorem 8.3 useless for the purposes of characterisational representationism, though they do certainly limit the kinds of accounts that might be built upon it. Certainly, we should say that an agent S

[98] **GRS1** would also need to be altered so as to include these more complex gambles in $\mathcal{G}$.

has credences $\mathcal{B}el$ and utilities $\mathcal{D}es$ *if and only if* her preferences over $\mathcal{G}$ satisfies Indifference between Obvious Equivalents, **GRS1**–**11** and Condition 1 in the appropriate way. The foregoing discussion suggests that satisfying those conditions should be considered neither necessary nor sufficient for having such-and-such credences and utilities. But we have already seen reasons for rejecting accounts along these lines (§3.3). Moreover, as we noted in §4.1, we should not expect—and it's no commitment of characterisational representationism—that a single representation theorem should do *all* the definitional heavy lifting when it comes to characterising credences and utilities. We can, and in some cases *should*, appeal to information which goes beyond just the agent's preferences when seeking to determine her credences and utilities; *a fortiori*, we can and should go beyond her preferences with regards a restricted space of objects $\mathcal{G}$.

The point of characterisational representationism should not be to show how credences and utilities simply *reduce* to a set of preference states and nothing more. They don't, so that project is a dead end. Instead, the most plausible *interpretivist* and *functionalist* approaches to the graded attitudes place a strong emphasis on their connection to preferences, without that connection being *all there is* to the possession of credences and utilities.

Let me put a bit more flesh on these bones. We know that ordinary agents don't evaluate gambles according to the Naïve Expected Utility formula. Suppose instead that *when they are fully rational*—i.e., they've thoughtfully considered all the possibilities, worked out all the logical relationships, and are free from any interfering influences (intoxication, sleep-deprivation, etc.)—then the value that they attach to an arbitrary *n*-outcome possible gamble is given by the standard expected utility formula:

$$\mathcal{D}es((o_i, P_i; \ldots; o_n, P_n)) = \mathcal{D}es(o_i \,\&\, P_i).\mathcal{B}el(P_i) + \ldots + \mathcal{D}es(o_n \,\&\, P_n).\mathcal{B}el(P_n)$$

But ordinary agents aren't fully rational, so we should not expect to be able to *T*-represent *S*'s preferences over the space of all possible gambles $\mathcal{G}^+$ so that they always come out as maximising expected utility according to the standard formula relative to some $\mathcal{B}el$ and $\mathcal{D}es$—and if it turns out that we can, then we have good reasons to think that $\mathcal{B}el$ and $\mathcal{D}es$ don't accurately model her credences and utilities. In short, the determinants of agents' preference patterns are complicated—they are strongly tied to credences and utilities, and perhaps more besides, but in ordinary circumstances such connections need not be rigidly systematic.

Suppose, however, that *under special conditions*, agents reliably have preferences in conformity with the standard formula. These are conditions where the agent is free from interfering influences, where the objects of preference aren't particularly complicated, and—most importantly—where the relevant logical relationships are all *obvious*. We might then use *S*'s restricted preferences in these special conditions to solve the problem of separability (§3.1) and help fix upon the contents of her attitudes—at least where those preferences can be associated with just one $\mathcal{B}el$ and $\mathcal{D}es$ assignment consistent with the hypothesis that she maximises expected utility. For this, a representation theorem would prove useful.

Thus, I think the best way to justify Theorem 8.3's emphasis on preferences over the simple two-outcome gambles in $\mathcal{G}$ is that these seem to be the kinds of preferences which are most likely to be maximally revealing *vis-à-vis* the agent's credences and utilities. The gambles in $\mathcal{G}$ are, in Ramsey's words, "the sorts of cases with which we are most concerned", where something like expected utility theory is most likely to be descriptively accurate, where a subject's credences and utilities are most likely to shine through in her preferences. We should not commit ourselves to the hypothesis that ordinary agents are expected utility maximisers *generally*, or even very often—in which case, our best bet is to narrow our focus to those circumstances where expected utility theory is more likely to be correct.

This is *not* to say that her preferences over other types of gambles—or any other intuitively relevant data, for that matter—should be *ignored* when trying to assign appropriate $\mathcal{B}el$ and $\mathcal{D}es$ functions to the agent—only that an interpretational priority might be given to these rather more straightforward gambles. Theorem 8.3, then, should not be taken to give us the *whole* story about an agent's credences and utilities and their functional role in relation to her preferences: having such-and-such credences and utilities is *not* simply a matter of having preferences which satisfy the stated conditions. But, from the perspective of deciding upon the *contents* of those states, it could be said to form a very important part.

CHAPTER NINE

Naturalisation and Characterisational Representationism

The goal of this work was to evaluate the status of characterisational representationism. There were two main questions to address. The first was whether, *given* the kinds of representation thoerems presently on offer, characterisational representationism could help us to directly advance the *naturalisation project* by connecting credences and utilities to the non-intentional world. The second was whether characterisational representationism, in any form, is a viable response to the *characterisation project*—whether, in particular, there is any point to developing representation theorems with the goal of understanding what it is to have credences and utilities in mind.

The Decision-theoretic Interpretation of a representation theorem T tells us that if an agent S's preferences over some collection of *basic objects of preference* ($\mathcal{BOP}$) satisfies a particular set of preference conditions C, then S can be represented as following some decision rule $\mathcal{R}$ with credences $\mathcal{B}el$ and utilities $\mathcal{D}es$. In Chapters 3 and 4, I argued that some such theorem T could be the basis for a plausible version of characterisational representationism, *if* it had the appropriate properties. The issue, then, was whether any such theorem existed.

Chapters 5 to 7 surveyed the majority of theorems presently on offer, and found that each came up short. Broadly put, there were five basic kinds of issues that were raised, clustered around the following themes:

1. *Satisfiability*: whether T's preference conditions C (under a reasonable interpretation) are typically satisfied (or approximately satisfied) by ordinary agents.
2. *Plausibility*: whether, under the condition that S satisfies C, the resulting representation of S's credences ($\mathcal{B}el$), utilities ($\mathcal{D}es$), and decision-making procedure ($\mathcal{R}$) is intuitively and empirically plausible.
3. *Uniqueness*: whether the resulting representation is, in an interesting sense, at least somewhat unique.
4. *Circularity*: whether any useful Decision-theoretic Interpretation of T (i.e., an interpretation of $\succcurlyeq$ on $\mathcal{BOP}$) depends on a prior specification of S's credences and utilities.

5. *Naturalisability*: whether any useful Decision-theoretic Interpretation of T involves an unavoidable appeal to some intentional state or other.

Every contemporary representation theorem raised issues of at least one of these kinds, and most raised issues of several kinds.

With respect to conditions the *Satisfiability* and *Plausibility* constraints, there are frequently expressed concerns that the preference conditions and expected utility models associated with CEU theorems in particular are descriptively implausible on the basis of decades of empirical research (§3.3.2). Those preference conditions are *perhaps* more plausible for ideally rational agents, but ordinary agents do not seem to satisfy them—at least not exactly. Likewise, while it may be plausible that ideally rational agents are probabilistically coherent expected utility maximisers, this is far less likely for ordinary agents (for whom psychologists have developed a wealth of more empirically successful models).

There are also frequently expressed concerns regarding the appropriate interpretation of the uniqueness results that attach to standard CEU theorems. However, as noted in §3.2, these concerns apply primarily to a very Naïve version of characterisational representationism. The real issue, if there is one, is justifying the appeal to a specific *representation scheme* (given by a theorem with sufficiently strong uniqueness conditions). Given both the intuitive appeal of expected utility maximisation and the fact that most current models of decision-making involve it or something much like it (§3.3.2), this does not seem to be a particularly pressing challenge—at least not when placed in comparison with the other issues that face the representation theorems we have now.

My critical discussion tended to generalise away from the issues that face CEU theorems specifically, and focused instead on concerns that arise for CEU and NCU theorems alike. For theorems developed in the Savage framework, given an interpretation of Savage's act-functions as representing *acts*, we are faced with the constant acts problem, which severely curtails the satisfiability of Savage's preference conditions by *any* agent (§5.2). A similar problem applies to any Savage-like theorem which requires a similarly rich space of act-functions (which seems to be all of them). Other interpretations of act-functions present their own, distinctive challenges—particularly regarding uniqueness (§5.2.4). The $\mathcal{B}el$ functions associated with Savagean theorems are also of limited empirical and intuitive plausibility, being (a) in most cases restricted to highly structured credence functions, such as probability functions and capacities (§5.5), and (b) unable to represent credences towards propositions about acts and (moreover) *any* proposition specifying something which is of importance to us which might depend on our acts (§5.3).[99]

[99] Savagean theorems are also, for that matter, generally incapable of representing utilities towards anything other than outcomes—i.e., propositions which are maximally specific with respect to what the agent cares about.

Finally, it appears to be impossible to specify an adequate interpretation of any Savage-like theorem without some prior access to subject's doxastic states (§5.4).

Theorems within the Anscombe and Aumann paradigm present essentially the same difficulties as Savagean theorems, and more besides (§6.1.2). In particular, it's unlikely that ordinary agents even *have* preferences over lotteries upon lotteries upon lotteries, and it's even more unlikely still that such preferences would play much of a role in fixing subjects' credences and utilities. Furthermore, the essential appeal to 'objective lotteries' implies that Anscombe and Aumann's theorem (and any other lottery-based theorem) cannot be given an adequate interpretation that is independent of substantive (and empirically dubious) background presuppositions about ordinary agents' credences.

Theorems which, like Ramsey's, rely essentially on ethically neutral propositions, present their own unique difficulties (§7.2). To the extent that there *are* no ethically neutral propositions, not even to a reasonable approximation, the preference conditions which mention them cannot be (non-trivially) satisfied. Ramsey's assumption that we have well-defined preferences over maximally (or near-maximally) specific propositions, and gambles involving such propositions as outcomes, is also highly problematic: the basic objects of preference that Ramsey considers seem too specific to even entertain.

Because Jeffrey appeals directly to mentalistic preferences rather than attempting to define $\mathcal{Bel}$ and $\mathcal{Des}$ in terms of choice dispositions, it is possible to specify what it would take for S to satisfy his theorem's preference conditions without prior access to S's doxastic states—though the use of an unreduced intentional notion does raise questions regarding the *Naturalisability* of such theorems (which will be discussed shortly). On the other hand, and partly because they have been developed with explicitly normative goals in mind, the (very few) monoset theorems which presently exist show substantial room for improvement *vis-à-vis* characterisational representationism (§6.2.2). In particular, they (i) place very strong restrictions on preferences, which only seem plausible for ideally rational agents; (ii) are limited to probabilistic $\mathcal{Bel}$ functions, which in most cases ranges over an infinite domain of ever-increasingly specific propositions; and (iii) are restricted to representing agents as expected utility maximisers across the board.

Theorem 8.3 is, in very broad terms, an amalgamation of Jeffrey's and Ramsey's ideas, with some unique features of its own. Ontologically, it is similar to Jeffrey's, while formally it has more in common with Ramsey's. It was developed to make some headway towards avoiding the *Satisfiability*, *Plausibility*, *Uniqueness* and *Circularity* issues that were raised for earlier theorems. Of particular note is the fact that Theorem 8.3's $\mathcal{Bel}$ and $\mathcal{Des}$ functions seem particularly well-suited for the representation of non-ideal agents' credences and utilities, especially in comparison to any of the other theorems discussed in previous chapters. The relevant points were summarised in §8.3.2, and I will not repeat them here. I have also argued that the posited decision rule is plausible, *given* the restrictions imposed on $\mathcal{G}$ and the assumed interpretation of $\succcurlyeq$ (§8.1.1).[100]

Furthermore, Theorem 8.3 also has the Standard Uniqueness Condition, and, as with the monoset theorems, there is no obvious appeal to agents' credences and/or utilities involved in the interpretation of any of the theorem's primitives. Finally, as I've argued in §8.2.2 and §8.3.3, Theorem 8.3 (or some weakening thereof) appears to have descriptively reasonable preference conditions—though whether this appearance is accurate is a matter for future empirical investigation. To the extent that the conditions do not seem reasonable—as, for example, with the rather idealised conditions Condition 1 and **GRS9**—there are still interesting (albeit weaker) representation results which might be established in their absence.

Theorem 8.3 does not present a solution to all of the technical problems facing characterisational representationism—it has only a very limited domain of application, and requires still some substantial degree of idealisation within that domain—but it does at least suggest that progress can be made towards improving the satisfiability of representation theorems' preference conditions to ordinary agents and the plausibility of the resulting representations. If what I have argued in Chapters 3 and 4 is right, then, there is motivation to continue developing representation theorems aimed at helping us to characterise the credences and utilities of ordinary agents—as a response to the characterisation project at least, there is promise in pursuing characterisational representationism.

A question remains regarding the naturalisation project. None of the theorems discussed satisfied the *Naturalisability* constraint. Savage's, Anscombe and Aumann's, and Ramsey's theorems exemplify three distinct kinds of framework built around (but not necessarily committed to) a behavioural conception of preference, and so offered the best hope for naturalisation. In each, the basic objects of preference are generally interpreted as *objects of choice*—i.e., *acts*, *lotteries*, and *gambles*, respectively. In §2.2, I raised some problems for the behavioural interpretation of $\succcurlyeq$, but even setting those problems aside, I have argued that no current representation theorem lends itself to a plausible and naturalistic interpretation suitable for the purposes of characterisational representationism. The basic reason for this was raised in §5.4.2 in relation to Savage's theorem, but we can see now that the point generalises easily.

In order to derive some unique (or even semi-unique) $\mathcal{Bel}$ and $\mathcal{Des}$ from preferences over some collection of objects of choice—whether they be *acts*, *lotteries*, or *gambles*—those objects must be modelled as having a certain kind of *structure*, which connects each

[100] Because Jeffrey's theorem does not involve a similar restriction on the domain of its $\succcurlyeq$—which is simply the set of all propositions towards which the agent has credences and utilities—agents who satisfy Jeffrey's conditions are therefore represented as $\mathcal{EU}$-maximisers *across the board*, rather than as $\mathcal{EU}$-maximisers with respect to a limited domain of choice.

one in a unique way to the objects in the domains of $\mathcal{B}el$ and $\mathcal{D}es$. For instance, Savage's act-functions are simply functions from states to outcomes; while Ramsey's gambles are just pairings of outcomes with pairs of complementary objects of uncertainty. It is these connections which are drawn upon to derive $\mathcal{B}el$ and $\mathcal{D}es$—without them, any pattern of behavioural preferences could be correlated with any set of credences and utilities we like. In all cases, then, the interpretive question arises: should this structure be taken to represent the *actual* properties of the relevant object of choice, or the properties that the decision-maker *thinks* are associated with the options available to her. If we suppose the former, then the theorem's $\mathcal{B}el$ and $\mathcal{D}es$ are all but guaranteed to be misrepresent the subject's actual credences and utilities. If we suppose the latter, however, then we have already given up on a naturalistic interpretation of the theorem in question.

There is no easy way around this problem, at least given anything like the theorems presently on offer. One potential response would be to build in to one's account a number of assumptions about how agents conceptualise their decision situations, which would have to be well-motivated and independently plausible. In §5.4.2, I argued that we *might* be able to take this strategy for *ideally rational* agents, for whom it may be somewhat plausible to assume that each act's actual causal profile is accurately represented using a set of highly specific *dependency hypotheses* as states. But ideally rational agents are vastly unlike ordinary agents in the relevant respects, and the latter are likely to vary in how they conceptualise their decision situations in highly non-systematic ways. Likewise, in applying Ramsey's system, we might assume that were a subject to be offered a collection of gambles, (i) she would fully understand what she is offered, (ii) would be certain that the payouts for whatever gamble she accepts will be exactly as described, and that (iii) the very presence of the offers would not seriously alter her attitudes. Individually, each assumption is dubious; in conjunction, they are almost certainly false.

What remains to be seen is whether mentalistic preferences can be naturalised. I do not have a full answer to this question, but I do want to say a few words about how I think naturalisation will *not* be achieved—namely, *via* their supposedly direct connections with behaviour. The problem, of course, is that only a very limited number of mentalistic preference states plausibly show any direct connections with behavioural patterns. If the goal is to naturalise mentalistic preferences, then we will need to specify what it is for $\succcurlyeq$ to hold between arbitrary propositions—and in many cases it seems implausible that a preference for P over Q will be *directly* apparent in behaviour independent of assumptions regarding beliefs.

An example will be helpful here. In his (1990), Jeffrey suggested a behavioural operationalisation of his use of $\succcurlyeq$, given in terms of reactions to *news items*:

> To say that [P] is ranked higher than [Q] [in the agent's preference ranking] means that the agent would welcome the news that [P] is true more than he would the news that [Q] is true: [P] would be better news than [Q]. (82)

Such an operationalisation cannot underlie a fully naturalistic account of mentalistic preference, however, as we can only interpret an agent's reactions to news items if we know how she *understands* those items—and it does not seem plausible that we could have such knowledge without prior access to her doxastic states. (Furthermore, we had better hope that the agent hasn't, for whatever reason, decided to hide her true reactions in the hope of misleading us.)

There is, to be sure, a small subset of propositions such that, if S *were* to have preferences over them, S likely *would* have a particular pattern of behavioural dispositions. As Jeffrey describes them, these are the propositions which specify behaviours which the agent can assuredly make true by a pure exercise of the will. Let us call these *action-propositions*. It seems reasonable to suppose that a (mentalistic) preference for one action-proposition P over another Q would be directly manifest in behaviour: if S prefers P to Q and knows (i.e., with certainty) she can make P true by a pure exercise of the will α, and Q true by a pure exercise of the will β, then S should have a behavioural preference for α over β.

Of course, it should not be taken for granted that ordinary agents even *have* preferences over action-propositions—and even supposing that they do, it would be difficult to determine *which* propositions are action-propositions for S without first peaking inside her head. (One can presumably be mistaken about what things one can make true by a pure exercise of the will.) And, finally, it's still more difficult to see how S's mentalistic preferences over *non-action propositions* might be linked to her behaviour without the mediation of other mental states.

It is not clear, then whether mentalistic preferences are readily naturalisable—preferences probably don't reduce directly to behaviour, but that was perhaps the wrong place to look in any case. Moreover (as the foregoing makes clear) an attempt to cash out the mentalistic notion of preference in terms of behaviour seems to require some appeal to a background doxastic state, presenting the threat of vicious circularity. As Stalnaker put it upon raising an analogous concern for his own account (discussed in §4.4), "Is this theory simply a shell game that hides the problem of intentionality under belief [or something belief-like] while it explains desire [or something desire-like], and under desire while it explains belief?" (1984, 15).

As Stalnaker tries to do for his account of belief and desire, characterisational representationism will need its own way of breaking out of this circle—representation theorems (at least of the kind we have now) won't let us pin down subjects' contents through sufficient observation of their behavioural dispositions. I have already suggested that credences should also be understood in terms of their connections with evidence and reasoning; in this respect, the functional role semantics suggested in §4.5 is similar to Stalnaker's approach for circumventing the circularity issue. But I suspect that more will be needed: we should look for a characterisation of (mental) preferences, and one which is not (or not wholly) given in terms of their connection to behaviours. Where naturalisation is a key constraint, causal-informational or teleosemantic views may be of use here. Alternatively, one might forego the naturalisation project in favour of a non-reductive characterisation of credences and utilities (e.g., Schwitzgebel 2002, 2013), an approach which takes certain intentional states as basic (as Eriksson and Hájek 2007 suggest for credences), or an approach which explains the content of preferences via their phenomenological connections (cf. Pautz 2013).

It is in the end therefore unclear whether a version of characterisational representationism based on something like Theorem 8.3 or any ontologically similar theorem will help to directly advance the naturalisation project. I would, however, offer a more modest suggestion: whether we eventually find a way to naturalise mentalistic preferences or not, having an improved understanding of the connections between them and our credences and utilities certain won't *hurt*—and it is reasonable to suppose that a representation theorem with the right properties could be very useful in developing such an understanding.

Appendix A: Proofs

Theorem 8.1

The proof of Theorem 8.1 proceeds as follows. First, we show that **GRS1–7** jointly entail that $<\mathcal{O} \times \mathcal{O}, \geq^{d}>$ is an algebraic difference structure, allowing us to invoke Theorem 8.2 giving us $\mathcal{D}es$ on $\mathcal{O}$. **GRS8** and **GRS9** are then used to extend $\mathcal{D}es$ to $\mathcal{O} \cup \mathcal{G}$, and it's shown that this provides us with an interval scale representation of $\succcurlyeq$ on $\mathcal{O} \cup \mathcal{G}$.[101]

It will be helpful to establish three lemmas first:

Lemma A

For every pair $o_1, o_2 \in \mathcal{O}$, there is a $(o'_1, \pi; o'_2) \in \mathcal{G}$

1. Follows immediately from **GRS1** and **GRS2**. ■

We thus know that universally quantified statements about possible gambles conditional on some proposition of credence ½ are never trivially satisfied; so, for instance, where a step says 'for all $(o'_1, \pi; o'_4)$, $(o'_2, \pi'; o'_3) \in \mathcal{G}$, $(o'_1, \pi; o'_4) \succcurlyeq (o'_2, \pi'; o'_3)$', Lemma A ensures that at least one such pair of gambles exists in $\mathcal{G}$. I will generally omit this step in what follows. Set memberships have been suppressed where obvious: henceforth we are only concerned with gambles in $\mathcal{G}$.

Lemma B

If $(o'_1, \pi; o'_4) \succcurlyeq (o'_2, \pi'; o'_3)$ for some pair $(o'_1, \pi; o'_4)$, $(o'_2, \pi'; o'_3)$, then $(o_1, o_2) \geq^{d} (o_3, o_4)$

1. Suppose that $(o'_1, \pi; o'_4) \succcurlyeq (o'_2, \pi'; o'_3)$ for some such pair.
2. By Lemma A, some $(o''_4, \pi''; o''_1)$ exists, and by successive iterations of **GRS4**, $(o'_1, \pi; o'_4) \sim (o''_4, \pi''; o''_1)$ and $(o''_4, \pi''; o''_1) \sim (o'''_1, \pi'''; o'''_4)$ for all such pairs. Because $\sim$ is an equivalence relation (**GRS3**), $(o_1, \pi; o_4) \sim (o'''_1, \pi'''; o'''_4)$ for all such pairs.
3. By the same steps, we know that $(o'_2, \pi'; o'_3) \sim (o''_2, \pi^*; o''_3)$ for all such pairs.

[101] Several of the steps in what follows owe much to (Bradley 2001), especially Lemma C and the steps involving it.

4. So given our starting supposition, $(o'''_1, \pi'''; o'''_4) \succcurlyeq (o''_2, \pi^*; o''_3)$ for all such pairs, which is just the right hand side of Definition 8.2. ■

Lemma C

If $(o_1, o_2) \geq^{d} (o_3, o_4)$, then $(o_4, o_3) \geq^{d} (o_2, o_1)$, and $(o_1, o_3) \geq^{d} (o_2, o_4)$

1. Suppose $(o_1, o_2) \geq^{d} (o_3, o_4)$, so $(o'_1, \pi; o'_4) \succcurlyeq (o'_2, \pi'; o'_3)$ for all such gambles.
2. Lemma A ensures some $(o''_4, \pi^*; o''_1)$, $(o''_3, \pi^+; o''_2)$ exist, and by **GRS4**, $(o''_4, \pi^*; o''_1) \sim (o'_1, \pi; o'_4)$ and $(o''_3, \pi^+; o''_2) \sim (o'_2, \pi'; o'_3)$. Substituting for equally valued gambles, we get $(o''_4, \pi^*; o''_1) \succcurlyeq (o''_3, \pi^+; o''_2)$, which given Lemma B implies $(o_4, o_3) \geq^{d} (o_2, o_1)$.
3. Likewise, $(o'_1, \pi; o'_4) \succcurlyeq (o''_3, \pi^+; o''_2)$, so $(o_1, o_3) \geq^{d} (o_2, o_4)$. ■

We can now show that **ADS1–5** follow from **GRS1–7**. **ADS2** is simply the first part of Lemma C. Next we will prove that $\geq^{d}$ on $\mathcal{O} \times \mathcal{O}$ is complete:

1. From Lemma A, for any two (o_1, o_4), (o_2, o_3), there exist $(o'_1, \pi; o'_4)$, $(o'_2, \pi'; o'_3)$. From **GRS3**, either $(o'_1, \pi; o'_4) \succcurlyeq (o'_2, \pi'; o'_3)$ or $(o'_2, \pi'; o'_3) \succcurlyeq (o'_1, \pi; o'_4)$.
2. Given Lemma B, if the former then $(o_1, o_2) \geq^{d} (o_3, o_4)$, and if the latter then $(o_3, o_4) \geq^{d} (o_1, o_2)$. So either $(o_1, o_2) \geq^{d} (o_3, o_4)$ or $(o_3, o_4) \geq^{d} (o_1, o_2)$. ■

We also prove that $\geq^{d}$ on $\mathcal{O} \times \mathcal{O}$ is transitive:

1. Suppose that $(o_1, o_2) \geq^{d} (o_3, o_4)$ and $(o_3, o_4) \geq^{d} (o_5, o_6)$.
2. From Definition 8.2, for all the relevant gambles, this implies that $(o'_1, \pi; o'_4) \succcurlyeq (o'_2, \pi'; o'_3)$ and $(o'_3, \pi^*; o'_6) \succcurlyeq (o'_4, \pi^+; o'_5)$.
3. For any pair of gambles $(o''_1, \pi''; o''_6)$, $(o''_2, \pi'''; o''_5)$, **GRS5** then requires that $(o''_1, \pi''; o''_6) \succcurlyeq (o''_2, \pi'''; o''_5)$, and $(o_1, o_2) \geq^{d} (o_5, o_6)$ follows from Lemma B. ■

So $\geq^{d}$ on $\mathcal{O} \times \mathcal{O}$ is a weak ordering and **ADS1** is satisfied. Next we show that **ADS3** is satisfied:

1. Suppose $(o_1, o_2) \geq^{d} (o_4, o_5)$ and $(o_2, o_3) \geq^{d} (o_5, o_6)$.
2. The second part of Lemma C applied to each conjunct entails $(o_1, o_4) \geq^{d} (o_2, o_5)$ and $(o_2, o_5) \geq^{d} (o_3, o_6)$. Because $\geq^{d}$ is transitive, $(o_1, o_4) \geq^{d} (o_3, o_6)$. So from Lemma C again, $(o_1, o_3) \geq^{d} (o_4, o_6)$. ■

ADS4 is satisfied:

1. Suppose $(o_1, o_2) \geq^{d} (o_3, o_4) \geq^{d} (o_1, o_1)$.

2. From **GRS6**, for every triple o_1, o_3, o_4, there is a o_5 such that for some $(o'_1, \pi; o'_4)$, $(o_5, \pi'; o'_3)$ (ensured by Lemma A), $(o'_1, \pi; o'_4) \sim (o_5, \pi'; o'_3)$. Applying Lemma B, we see that there must be a o_5 such that $(o_1, o_5) =^d (o_3, o_4)$.
3. Likewise, for every triple o_3, o_4, o_2, there is a o_6 such that $(o'_3, \pi; o'_2) \sim (o_6, \pi'; o'_4)$ for some such pair; so there is a o_6 such that $(o_3, o_4) =^d (o_6, o_2)$. ■

And **ADS5** is also satisfied. The proof of this is trivial given **GRS7**, Definition 8.2, and the definition of a strictly bounded standard sequence; it has therefore been left unstated. **GRS1**–**7** therefore imply that $<\mathcal{O} \times \mathcal{O}, \geq^d>$ is an algebraic difference structure, which ensures the existence of the appropriate $\mathcal{D}es$ on $\mathcal{O}$ (unique up to positive linear transformation), such that:

$$(o_1, o_2) \geq^d (o_3, o_4) \text{ iff } \mathcal{D}es(o_1) - \mathcal{D}es(o_2) \geq \mathcal{D}es(o_3) - \mathcal{D}es(o_4)$$

We appeal primarily to **GRS8** to show that $\mathcal{D}es(o_1) \geq \mathcal{D}es(o_2)$ iff $o_1 \succcurlyeq o_2$:

1. From **GRS8**, $o'_1 \sim o_1$ iff, for all $(o''_1, P; o'''_1)$, $o_1 \sim (o''_1, P; o'''_1)$; and similarly, $o'_2 \sim o_2$ iff, for all $(o''_2, P; o'''_2)$, $o_2 \sim (o''_2, P; o'''_2)$.
2. Given **GRS3** then, $o_1 \succcurlyeq o_2$ iff $(o''_1, \pi; o'''_1) \succcurlyeq (o''_2, \pi'; o'''_2)$ for all such gambles, which holds iff $(o_1, o_2) \geq^d (o_2, o_1)$.
3. From Theorem 8.2, $(o_1, o_2) \geq^d (o_2, o_1)$ iff $\mathcal{D}es(o_1) - \mathcal{D}es(o_2) \geq \mathcal{D}es(o_2) - \mathcal{D}es(o_1)$, which can only be if $\mathcal{D}es(o_1) \geq \mathcal{D}es(o_2)$. So $o_1 \succcurlyeq o_2$ iff $\mathcal{D}es(o_1) \geq \mathcal{D}es(o_2)$. ■

We further require that $\mathcal{D}es$ is defined on $\mathcal{O} \cup \mathcal{G}$. From **GRS9**, we know that for every $(o_1, P; o_2)$ there is a o_3 such that $(o_1, P; o_2) \sim o_3$. We can achieve the desired extension by making the following stipulation:

$$\text{For all } o_3, (o_1, P; o_2) \in \mathcal{O} \cup \mathcal{G}, \mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es(o_3) \text{ iff } (o_1, P; o_2) \sim o_3$$

The proof that condition (i) of Theorem 8.1 then holds is trivial and left unstated. The uniqueness properties of $\mathcal{D}es$ on $\mathcal{O}$ will also clearly hold for $\mathcal{D}es$ on $\mathcal{O} \cup \mathcal{G}$. The foregoing thus establishes Theorem 8.1.

Theorem 8.3

To prove Theorem 8.3, we need to show that Definition 8.7 provides us with a unique function $\mathcal{B}el$ that satisfies the stated properties. To begin with, I will prove that $\mathcal{B}el$ is a credence function defined for all $P \in \mathcal{P}$.

1. That for each $P \in \mathcal{P}$ we will always be able to find outcomes and gambles satisfying Definition 8.7's conditions follows immediately from **GRS1** and **GRS10**.
2. That $\mathcal{B}el(P)$ is independent of the choice of outcomes and gambles satisfying the antecedent conditions follows immediately from Condition 1.
3. The range of $\mathcal{B}el$ is [0, 1]: from **GRS8** and **GRS3**, for all $(o_1, P; o_2)$, either $o_1 \succcurlyeq o_2$ and $o_1 \succcurlyeq (o_1, P; o_2) \succcurlyeq o_2$, or $o_2 \succcurlyeq o_1$ and $o_2 \succcurlyeq (o_1, P; o_2) \succcurlyeq o_1$. Given the established properties of $\mathcal{D}es$, we know $\mathcal{D}es((o_1, P; o_2))$ always sits somewhere weakly between $\mathcal{D}es(o_1)$ and $\mathcal{D}es(o_2)$. It follows that the ratio of the difference between $\mathcal{D}es((o_1, P; o_2))$ and $\mathcal{D}es(o_2)$ and the difference between $\mathcal{D}es(o_1)$ and $\mathcal{D}es(o_2)$ will always be within [0, 1]. ■

We can now prove condition (iii), i.e., $\mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es(o_1).\mathcal{B}el(P) + \mathcal{D}es(o_2). (1 - \mathcal{B}el(P))$:

1. Suppose first that $o_1 \sim o_2$; then, by reasoning noted above, $\mathcal{D}es(o_1) = \mathcal{D}es(o_2) = \mathcal{D}es((o_1, P; o_2))$. Let $\mathcal{D}es(o_1) = x$. The required equality then holds iff $x = x.\mathcal{B}el(P) + x.(1 - \mathcal{B}el(P))$; we have already noted that $\mathcal{B}el(P) \in [0, 1]$, so this can be assumed regardless of the value of $\mathcal{B}el(P)$.
2. Suppose next that $\neg(o_1 \sim o_2)$. From Definition 8.7, $\mathcal{B}el(P) = (\mathcal{D}es((o_1, P; o_2)) - \mathcal{D}es(o_2)) / (\mathcal{D}es(o_1) - \mathcal{D}es(o_2))$, which holds iff
   $(\mathcal{D}es(o_1) - \mathcal{D}es(o_2)).\mathcal{B}el(P) = \mathcal{D}es((o_1, P; o_2)) - \mathcal{D}es(o_2)$ iff
   $\mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es(o_1).\mathcal{B}el(P) - \mathcal{D}es(o_2).\mathcal{B}el(P) + \mathcal{D}es(o_2)$ iff
   $\mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es(o_1).\mathcal{B}el(P) + \mathcal{D}es(o_2).(1 - \mathcal{B}el(P))$. ■

If **GRS11** holds, then condition (iv), that $\mathcal{B}el(P) = 1 - \mathcal{B}el(\neg P)$, holds.[102]

1. As already shown, for all $P \in \mathcal{P}$, there is some $(o_1, P; o_2)$ such that the ratio $(\mathcal{D}es((o_1, P; o_2)) - \mathcal{D}es(o_2)) / (\mathcal{D}es(o_1) - \mathcal{D}es(o_2))$ is defined (i.e. such that $\neg(o_1 \sim o_2)$); from the foregoing proofs, this ratio is the value of $\mathcal{B}el(P)$.
2. From **GRS1** and since $\mathcal{P}$ is closed under negation, if $(o_1, P; o_2)$ is in $\mathcal{G}$ then $(o_2, \neg P; o_1)$ is in $\mathcal{G}$; thus $\mathcal{B}el(\neg P) = (\mathcal{D}es((o_2, \neg P; o_1)) - \mathcal{D}es(o_1)) / (\mathcal{D}es(o_2) - \mathcal{D}es(o_1))$.
3. Multiplying the denominator and the numerator by –1 gets us $\mathcal{B}el(\neg P) = (\mathcal{D}es(o_1) - \mathcal{D}es((o_2, \neg P; o_1))) / (\mathcal{D}es(o_1) - \mathcal{D}es(o_2))$.
4. **GRS11** ensures $(o_1, P; o_2) \sim (o_2, \neg P; o_1)$, so $\mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es((o_2, \neg P; o_1))$.
5. Let $\mathcal{D}es((o_1, P; o_2)) = x$. Given the foregoing, $\mathcal{B}el(P) + \mathcal{B}el(\neg P)$ is equal to:
   $((x - \mathcal{D}es(o_2)) / (\mathcal{D}es(o_1) - \mathcal{D}es(o_2))) + ((\mathcal{D}es(o_1) - x) / (\mathcal{D}es(o_1) - \mathcal{D}es(o_2)))$
   $= (x - o_2 + o_1 - x) / (o_1 - o_2) = (x - x + o_1 - o_2) / (o_1 - o_2)$
   $= (o_1 - o_2) / (o_1 - o_2) = 1$. Condition (iv) follows immediately. ■

Finally, $\mathcal{B}el$ is unique:

[102] Thanks to Rachael Briggs for the main outline of the following proof.

1. From the earlier proofs and the fact that ratios of differences are preserved across admissible transformations of $\mathcal{D}es$, we know that there is only one function $\mathcal{B}el$ on $\boldsymbol{\mathcal{P}}$ such that $\mathcal{B}el(\boldsymbol{\mathcal{O}}) = 1$, $\mathcal{B}el(\emptyset) = 0$, and for any contingent P, if o_1, o_2 are such that $\neg(o_1 \sim o_2)$ and $(o_1, P; o_2)$, then $\mathcal{B}el(P) = \mathcal{D}es((o_1, P; o_2)) - \mathcal{D}es(o_2) / \mathcal{D}es(o_1) - \mathcal{D}es(o_2)$.
2. We have also already established that the previous equality holds iff $\mathcal{D}es((o_1, P; o_2)) = \mathcal{D}es(o_1).\mathcal{B}el(P) - \mathcal{D}es(o_2).\mathcal{B}el(P) + \mathcal{D}es(o_2)$. Since there is only one function satisfying the left-hand side, only one satisfies the right-hand side. ■

Appendix B: Varieties of Savagean Theorem

In this appendix, I will briefly outline four important examples of representation theorems developed within the Savage paradigm. The first is Luce and Krantz's *conditional expected utility* theory (a variation on classical expected utility theory), and the following three are recent important NCU theorems. Besides the theorems discussed here, other examples of NCU theorems within the Savage paradigm can be found in (Schmeidler 1989), (Wakker 1989), (Sarin and Wakker 1992), (Machina and Schmeidler 1992), (Casadesus-Masanell, Klibanoff *et al.* 2000), (Ghirardato, Maccheroni *et al.* 2003), and (Qu 2015).

Luce and Krantz (1971) suggest that a major problem with Savage's theorem is that his act-functions are defined for all states, which leads to the constant acts problem to be discussed below. Consequently, Luce and Krantz attempt to develop a representation theorem over a much more limited set of act-functions, which we will designate $\boldsymbol{\mathcal{A}}_{\mathbf{LK}}$. We begin with a set of states $\boldsymbol{\mathcal{S}}$ and outcomes $\boldsymbol{\mathcal{O}}$, however the states in $\boldsymbol{\mathcal{S}}$ need not be (and in general will not be) independent of all the available acts. $\boldsymbol{\mathcal{E}}$ is the set of all subsets of $\boldsymbol{\mathcal{S}}$, and a set $\boldsymbol{\mathcal{N}}$ of null events is defined. An act-function $\mathcal{F} \in \boldsymbol{\mathcal{A}}_{\mathbf{LK}}$ is then a function from a *non-null* event into $\boldsymbol{\mathcal{O}}$. Essentially, every $\mathcal{F} \in \boldsymbol{\mathcal{A}}_{\mathbf{LK}}$ is simply a restriction of one of Savage's total act-functions to some non-null event.

In order to ensure that $\succcurlyeq$ on $\boldsymbol{\mathcal{A}}_{\mathbf{LK}}$ has a rich enough structure to underlie their desired representation, Luce and Krantz need to assume the following three structural conditions:

If $E^* \subset E$ and $E^* \notin \boldsymbol{\mathcal{N}}$, then $\mathcal{F}_E \in \boldsymbol{\mathcal{A}}_{\mathbf{LK}}$ iff $\mathcal{F}_{E^*} \in \boldsymbol{\mathcal{A}}_{\mathbf{LK}}$
If $(E \cap E^*) = \emptyset$, then, if $\mathcal{F}_E, \mathcal{G}_{E^*} \in \boldsymbol{\mathcal{A}}_{\mathbf{LK}}$, then $\mathcal{F}_E \cup \mathcal{G}_{E^*} \in \boldsymbol{\mathcal{A}}_{\mathbf{LK}}$
For all $\mathcal{F}_E \in \boldsymbol{\mathcal{A}}_{\mathbf{LK}}$ and any $E^* \in \boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{N}}$, there is a $\mathcal{G}_{E^*} \in \boldsymbol{\mathcal{A}}_{\mathbf{LK}}$ such that $\mathcal{G}_{E^*} \sim \mathcal{F}_E$

Luce and Krantz are then able to prove that if $\succcurlyeq$ on $\boldsymbol{\mathcal{A}}_{\mathbf{LK}}$ satisfies their stated preference conditions (C_{LK}), then there will exist a finitely additive probability function $\mathcal{B}el$ on $\boldsymbol{\mathcal{E}}$ and a $\mathcal{D}es$: $\boldsymbol{\mathcal{A}}_{\mathbf{LK}} \mapsto \mathbb{R}$, such that:

(i) $E \in \boldsymbol{\mathcal{N}}$ iff $\mathcal{B}el(E) = 0$
(ii) $\mathcal{F}_E \succcurlyeq \mathcal{G}_{E^*}$ iff $\mathcal{D}es(\mathcal{F}_E) \geq \mathcal{D}es(\mathcal{G}_{E^*})$
(iii) If $(E \cap E^*) = \emptyset$, $\mathcal{D}es(\mathcal{F}_E \cup \mathcal{G}_{E^*}) = \mathcal{D}es(\mathcal{F}_E).\mathcal{B}el(E|E \cup E^*) + \mathcal{D}es(\mathcal{G}_{E^*}).\ \mathcal{B}el(E^*|E \cup E^*)$

Furthermore, they show that $\mathcal{B}el$ is unique and $\mathcal{D}es$ is unique up to positive linear transformation. They also note that while the domain of their $\mathcal{D}es$ function is $\mathcal{A}_{\text{LK}}$, it's possible to add two further preference conditions to their original theorem which allows for a $\mathcal{D}es$ function on $\mathcal{O}$ and a slightly different expected utility representation of $\succcurlyeq$ on a subset of $\mathcal{A}_{\text{LK}}$. Satisfaction of the two further conditions, however, requires readmitting constant act-functions into the space of act-functions over which $\succcurlyeq$ is defined.

Lara Buchak's (2013, 2014) theorem for *risk-weighted expected utility theory* also builds on essentially the same resources as Savage's theorem, with $\succcurlyeq$ being defined on the set of all finitely-valued act-functions in $\mathcal{O}^{\mathcal{S}}$. However, by setting different preference conditions C_{REU} on $\succcurlyeq$ than Savage does, she arrives at a wholly distinct form of representation that involves a probability function $\mathcal{B}el$, a utility function $\mathcal{D}es$, and a so-called *risk function*, $\mathcal{R}$, which is intended to reflect the degree to which an agent is risk averse. A function $\mathcal{R}$: [0, 1] ↦ [0, 1] is a risk function iff:

$$
\begin{aligned}
&\mathcal{R}(0) = 0\\
&\mathcal{R}(1) = 1\\
&\text{If } n \leq m, \text{ then } \mathcal{R}(n) \leq \mathcal{R}(m)\\
&\text{If } n < m, \text{ then } \mathcal{R}(n) < \mathcal{R}(m)
\end{aligned}
$$

Buchak is able to prove that if the conditions C_{REU} are satisfied, then there exists a finitely additive probability function $\mathcal{B}el$ on $\mathcal{E}$, a risk function $\mathcal{R}$, and a real-valued function $\mathcal{D}es$ on $\mathcal{O}$, that together determine a risk-weighted expected utility ($\mathcal{R}eu$) function to represent* $\succcurlyeq$ on all finite-valued act-functions in $\mathcal{A}$, where $\mathcal{R}eu$ is defined as follows: for every $\mathcal{F} \in \mathcal{A}$, $\mathcal{F} = (E_i, o_i \,|\, \ldots \,|\, E_{\text{n}}, o_{\text{n}})$,

$$
\begin{aligned}
\mathcal{R}eu(\mathcal{F}) = \;& \mathcal{D}es(o_1) + \mathcal{R}\big(\textstyle\sum_{i=2}^{n} \mathcal{B}el(E_{\text{i}})\big).(\mathcal{D}es(o_2) - \mathcal{D}es(o_1)) + \mathcal{R}\big(\textstyle\sum_{i=3}^{n} \mathcal{B}el(E_{\text{i}})\big).\,(\mathcal{D}es(o_3) -\\
& \mathcal{D}es(o_2)) + \ldots + \mathcal{R}(\mathcal{B}el(E_{\text{n}})).(\mathcal{D}es(o_{\text{n}}) - \mathcal{D}es(o_{\text{n}-1}))
\end{aligned}
$$

Given conditions C_{REU}, it then follows that

$$\mathcal{F} \succcurlyeq \mathcal{G} \text{ iff } \mathcal{R}eu(\mathcal{F}) \geq \mathcal{R}eu(\mathcal{G})$$

Furthermore, $\mathcal{R}$ is unique, while $\mathcal{B}el$ and $\mathcal{D}es$ have the Standard Uniqueness Condition. Buchak's theorem was developed with normative considerations in mind, hence the willingness to adopt a probabilistically coherent credence function $\mathcal{B}el$. The addition of a risk function is motivated by normative considerations, which suggest that CEU inadequately deals with rational attitudes towards risky prospects.

On the other hand, *cumulative prospect theory* (Tversky and Kahneman 1992, Wakker and Tversky 1993) was developed with explicitly descriptive aspirations, being primarily a response to the empirical evidence that ordinary agents often fail to be rational in various ways. This theory takes some work to spell out. Suppose, first of all, that a given element o_s of $\mathcal{O}$ can be designated as the *status quo*—that is, o_s is the outcome in which nothing of interest to the decision-maker changes. An outcome o is then considered *positive* iff the constant act-function for o is considered strictly preferable to the constant act-function for o_s; i.e., o is *positive* iff $\underline{o} \succ \underline{o}_s$. These outcomes are considered to be *gains* from the decision-maker's perspective. Likewise, o is *negative* iff $\underline{o}_s \succ \underline{o}$. These outcomes are then to be considered *losses*. As with Savage's act-functions, each act-function in cumulative prospect theory can be represented using the general form as a sequence of pairs of events and outcomes, but with one small notational difference: the outcomes should always be arranged from negative to positive, in increasing order. So, every act-function can be represented by $(E_{-m}, o_{-m} \mid \ldots \mid E_i, o_i \mid \ldots \mid E_n, o_n)$ where the set $\{E_{-m}, \ldots, E_i, \ldots, E_n\}$ is a partition of $\mathcal{S}$, and the outcomes o_{-m} to o_n are arranged such that o_i comes after o_j iff $\underline{o}_i \succ \underline{o}_j$. Let $\mathcal{F}(E_i) = \mathcal{F}(s)$, for any $s \in E_i$.

Next, for any act-function $\mathcal{F}$, we can define the *positive part of* $\mathcal{F}$, or $\mathcal{F}^+$, as follows:

$$\mathcal{F}^+(s) = \mathcal{F}(s) \text{ if } \mathcal{F}(s) \text{ is positive, and } o_s \text{ otherwise}$$

In other words, the positive part of $\mathcal{F}$ is an act-function $\mathcal{F}^+$ which is the same as $\mathcal{F}$ for all states s where $\mathcal{F}$ maps s to a positive outcome, but maps all other states to the status quo. The *negative part of* $\mathcal{F}$, or $\mathcal{F}^-$, is given a similar definition, *mutatis mutandis*. The purpose of dividing an act-function into its positive and negative parts is so that we can treat the valuation of the two parts differently. In particular, a representation theorem for cumulative prospect theory says that if its preference conditions (C_{CPT}) are satisfied by $\succcurlyeq$ on $\mathcal{A}$, then there will be a strictly increasing utility function $\mathcal{D}es$ satisfying $\mathcal{D}es(o_s) = 0$, unique up to a positive multiplicative constant, and *two* unique capacities $\mathcal{W}^+$ and $\mathcal{W}^-$, such that for $\mathcal{F} = (E_{-m}, o_{-m} \mid \ldots \mid E_i, o_i \mid \ldots \mid E_n, o_n)$, and $-m \leq i \leq n$,

$$\mathcal{C}pt(\mathcal{F}) = \mathcal{C}pt(\mathcal{F}^+) + \mathcal{C}pt(\mathcal{F}^-)$$

Where the two parts, $\mathcal{C}pt(\mathcal{F}^+)$ and $\mathcal{C}pt(\mathcal{F}^-)$, are defined as:

$$\mathcal{C}pt(\mathcal{F}^+) = \sum_{i=0}^{n} \pi_i^+ . \mathcal{D}es(\mathcal{F}(E_i)),$$

$$\mathcal{C}pt(\mathcal{F}^-) = \sum_{i=-m}^{0} \pi_i^- . \mathcal{D}es(\mathcal{F}(E_i))$$

And:

$$\mathcal{F} \succcurlyeq \mathcal{G} \text{ iff } \mathcal{C}pt(\mathcal{F}) \geq \mathcal{C}pt(\mathcal{G})$$

The so-called *decision-weights*, $\pi^+ = (\pi_0^+, \ldots, \pi_n^+)$ and $\pi^- = (\pi_{-m}^+, \ldots, \pi_0^+)$, are then defined:

$$\begin{aligned}
\pi_n^+ &= \mathcal{W}^+(E_n) \\
\pi_i^+ &= \mathcal{W}^+(E_i \cup \ldots \cup E_n) - \mathcal{W}^+(E_{i+1} \cup \ldots \cup E_n), \text{ for } 0 \leq i \leq n-1 \\
\pi_{-m}^- &= \mathcal{W}^-(E_{-m}) \\
\pi_i^- &= \mathcal{W}^-(E_{-m} \cup \ldots \cup E_i) - \mathcal{W}^-(E_{-m} \cup \ldots \cup E_{i-1}), \text{ for } 1-m \leq i \leq 0
\end{aligned}$$

If it's now supposed that $\pi_i = \pi_i^+$ whenever $i \geq 0$, and $\pi_i = \pi_i^-$ whenever $i < 0$, then $\mathcal{C}pt$ can be simplified to:

$$\mathcal{C}pt(\mathcal{F}) = \textstyle\sum_{i=-m}^{n} \pi_i . \mathcal{D}es(\mathcal{F}(E_i))$$

Kahneman and Tversky are quick to point out that their ‘decision weight’, π, is not to be interpreted as a representation of credences:

> Consider a gamble in which one can win 1,000 [dollars] or nothing, depending on the toss of a fair coin. For any reasonable person, the probability of winning is .50 in this situation […] however, the decision weight […] which is derived from [his] choices is likely to be smaller than ½. Decision weights measure the impact of events on the desirability of prospects [i.e., acts], and not merely the perceived likelihood of these events. (1979, 280)

It has often been supposed that decision weights represent the composition of two distinct psychological factors: the agent's credences and the agent's attitudes towards decision-making in general—such as their attitudes towards risk and loss. See, for example, (Fellner 1961), (Tversky and Fox 1995), (Fox, Rogers *et al.* 1996), (Fox and Tversky 1998), (Gonzalez and Wu 1999), (Kilka and Weber 2001), and (Abdellaoui, Vossmann *et al.* 2005). Wakker (2004) attempts a decomposition of these decision weights into these two factors “based solely on observable [i.e., behavioural] preference” (236).

The final example of an NCU theorem within the Savage paradigm is the recent *maxmin expected utility* theorem of Alon and Schmeidler (2014). Maxmin theories tell us that in cases of uncertainty, the preferred option is (or should be) the option(s) with the best worst potential outcome—thus, by selecting that option, the agent guarantees that if even if that choice results in its least valuable outcome obtaining, that outcome is still at least as good as the worst outcome of any of the other available options. Alon and Schmeidler prove that if their conditions C_{MEU} are satisfied by $\succcurlyeq$ on the set of all finitely valued act-functions in $\boldsymbol{\mathcal{A}}$, then there will exist a continuous utility function $\mathcal{D}es$ on $\boldsymbol{\mathcal{O}}$, and a non-empty, closed and convex set $\boldsymbol{\mathcal{B}}$ of probability functions $\mathcal{P}r$ defined on $\boldsymbol{\mathcal{E}}$, such that:

$$\mathcal{F} \succcurlyeq \mathcal{G} \text{ iff } \min_{Pr\in \boldsymbol{B}} \int_S \mathcal{D}es(\mathcal{F}(.))\, d\,\mathcal{P}r \geq \min_{Pr\in \boldsymbol{B}} \int_S \mathcal{D}es(\mathcal{G}(.))\, d\,\mathcal{P}r$$

They also show that $\mathcal{D}es$ is unique up to positive linear transformation, the set $\boldsymbol{\mathcal{B}}$ is unique, and for some event $E \in \boldsymbol{\mathcal{E}}$,

$$0 < \min_{Pr\in \boldsymbol{B}} \mathcal{P}r(E) < 1$$

Bibliography

Abdellaoui, M., Vossmann, F. and Weber, M. (2005). 'Choice-Based Elicitation and Decomposition of Decision Weights for Gains and Losses Under Uncertainty.' *Management Science* **51** (9): 1384-99.

Adams, E. W. (1975). *The Logic of Conditionals*. Reidel.

Ahmed, A. (2014). *Evidence, Decision, and Causality*. Cambridge University Press.

Ahn, D. S. (2008). 'Ambiguity Without a State Space.' *Review of Economic Studies* **75** (1): 3-28.

Allais, M. (1953). 'Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Américaine.' *Econometrica* **21** (4): 503-46.

Alon, S. and Schmeidler, D. (2014). 'Purely subjective maxmin expected utility.' *Journal of Economic Theory* **152**: 382–412.

Anscombe, F. J. and Aumann, R. J. (1963). 'A Definition of Subjective Probability.' *The Annals of Mathematical Statistics* **34** (2): 199-205.

Anscombe, G. E. M. (1963). *Intention*. Blackwell.

Armendt, B. (1986). 'A foundation for causal decision theory.' *Topoi* **5** (1): 3-19.

Baker, L. R. (1995). *Explaining attitudes*. Cambridge University Press.

Balch, M. (1974). 'On recent developments in subjective expected utility'. In *Essays on Economic Behavior under Uncertainty*. M. Balch, D. McFadden and S. Wu, Eds. North-Holland Publishing Company: 45-54.

Balch, M. and Fishburn, P. C. (1974). 'Subjective Expected Utility for Conditional Primitives'. In *Essays on Economic Behavior under Uncertainty*. M. Balch, D. McFadden and S. Wu, Eds. North-Holland Publishing Company: 57-69.

Bartha, P. (2007). 'Taking stock of infinite value: Pascal's Wager and relative utilities.' *Synthese* **154** (1): 5-52.

Bernoulli, D. (1738). 'Specimen theoriae novae de mensura sortis.' *Commentarii academiae scientiarum imperialis Petropolitanae* **5**: 175-92.

Birnbaum, M. H. and Beeghley, D. (1997). 'Violations of branch independence in judgments of the value of gambles.' *Psychological Science* **8** (2): 87-94.

Birnbaum, M. H. and Chavez, A. (1997). 'Tests of theories of decision making: Violations of branch independence and distribution independence.' *Organizational Behavior and Human Decision Processes* **71** (2): 161-94.

Bjerring, J. C. (2013). 'Impossible Worlds and Logical Omniscience: an Impossibility Result.' *Synthese* **190**: 2505-24.

Blavatskyy, P. (2013). 'A Simple Behavioral Characterization of Subjective Expected Utility.' *Operations Research* **61** (4): 932-40.

Block, N. (1986). 'Advertisement for a Semantics for Psychology.' *Midwest Studies in Philosophy* **10** (1): 615-78.

Blume, L. E., Brandenburger, A. and Dekel, E. (1991). 'Lexicographic Probabilities and Choice Under Uncertainty.' *Econometrica* **59** (1): 61-79.

Blumson, B. (2012). 'Mental Maps.' *Philosophy and Phenomenological Research* **85** (2): 413-34.

Boghossian, P. A. (1993). 'Does an Inferential Role Semantics Rest upon a Mistake?' *Philosophical Issues* **3**: 73-88.

Bolker, E. D. (1966). 'Functions resembling quotients of measures.' *Transactions of the American Mathematical Society* **124** (2): 292-312.

—— (1967). 'A simultaneous axiomatization of utility and subjective probability.' *Philosophy of Science* **34** (4): 333-40.

Braddon-Mitchell, D. and Jackson, F. (1996). *Philosophy of Mind and Cognition*. Blackwell.

Bradley, R. (1998). 'A Representation Theorem for a Decision Theory with Conditionals.' *Synthese* **116** (2): 187-229.

—— (2001). 'Ramsey and the Measurement of Belief'. In *Foundations of Bayesianism*. D. Corfield and J. Williamson, Eds. Kluwer Academic Publishers: 263-90.

—— (2007). 'A unified Bayesian decision theory.' *Theory and Decision* **63** (3): 233-63.

Bradley, R. and Stefansson, H. O. (forthcoming). 'Counterfactual Desirability.' *British Journal for the Philosophy of Science*.

Braithwaite, R. B. (1946). 'Belief and Action.' *Proceedings of the Aristotelian Society, Supplementary Volumes* **20**: 1-19.

Bratman, M. (1987). *Intentions, plans, and practical reason*. Center for the Study of Language and Information.

Briggs, R. (2009). 'Distorted Reflection.' *Philosophical Review* **118** (1): 59-85.

Broome, J. (1991). *Weighing goods: Equality, uncertainty and time*. Basil Blackwell Press.

—— (1993). 'Can a Humean be moderate'. In *Value, Welfare and Morality*. R. G. Frey and C. W. Morris, Eds. Cambridge University Press: 51-73.

Buchak, L. (2013). *Risk and Rationality*. Oxford University Press.

—— (2014). 'Risk and Tradeoffs.' *Erkenntnis* **79** (6 Supplement): 1091-117.

Camerer, C. (1995). 'Individual Decision Making'. In *The Handbook of Experimental Economics*. J. H. Kagel and A. E. Roth, Eds. Princeton University Press: 587-703.

Camp, E. (2007). 'Thinking with Maps.' *Philosophical Perspectives* **21**: 145-82.

—— (2015). 'Concepts and Characterizations'. In *The Conceptual Mind: New Directions in the Study of Concepts*. E. Margolis and S. Laurence, Eds. MIT Press.

Casadesus-Masanell, R., Klibanoff, P. and Ozdenoren, E. (2000). 'Maxmin Expected Utility over Savage Acts with a Set of Priors.' *Journal of Economic Theory* **92**: 35-65.

Chalmers, D. (2011). 'Frege's Puzzle and the Objects of Credence.' *Mind* **120** (479): 587-635.

Christensen, D. (1996). 'Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers.' *The Journal of Philosophy* **93** (9): 450-79.

—— (2001). 'Preference-based arguments for probabilism.' *Philosophy of Science* **68** (3): 356-76.

—— (2004). *Putting Logic in its Place: Formal Constraints on Rational Belief*. Oxford University Press.

Churchland, P. (1981). 'Eliminative Materialism and the Propositional Attitudes.' *The Journal of Philosophy* **78**: 67-90.

Clarke, R. (2013). 'Belief is Credence One (In Context).' *Philosopher's Imprint* **13** (11): 1-18.

Davidson, D. (1973). 'Radical interpretation.' *Dialectica* **27** (3-4): 313-28.

—— (1980). 'Toward a Unified Theory of Meaning and Action.' *Grazer Philosophische Studien* **11**: 1-12.

—— (1990). 'The Structure and Content of Truth.' *The Journal of Philosophy* **87** (6): 279-328.

—— (1991). 'What Is Present to the Mind?' *Philosophical Issues* **1**: 197-213.

—— (2004). 'Expressing Evaluations'. In *Problems of Rationality* Oxford University Press: 19-38.

Davidson, D. and Suppes, P. (1956). 'A finitistic axiomatization of subjective probability and utility.' *Econometrica* **24** (3): 264-75.

Davidson, D., Suppes, P. and Siegel, S. (1957). *Decision making; an experimental approach*. Stanford University Press.

de Finetti, B. (1931). 'Sul Significato Soggettivo Della Probabilita.' *Fundamenta Mathematicae* **17** (1): 298-329.

—— (1964). 'Foresight: its logical laws in subjective sources'. In *Breakthroughs in Statistics*. S. Kotz and N. L. Johnson, Eds. Springer: 134-74.

—— (1974). *Theory of Probability*. John Wiley & Sons.

Debreu, G. (1959). 'Cardinal utility for even-chance mixtures of pairs of sure prospects.' *The Review of Economic Studies* **28** (3): 174-7.

Dempster, A. P. (1968). 'A Generalization of Bayesian Inference.' *Journal of the Royal Statistical Society Series B (Methodological)* **30**: 205-47.

Dennett, D. C. (1971). 'Intentional Systems.' *The Journal of Philosophy* **68** (4): 87-106.

—— (1989). *The Intentional Stance*. MIT Press.

—— (1991). 'Real Patterns.' *The Journal of Philosophy* **88** (1): 27-51.

Dietrich, F. and List, C. (2013). 'Where do preferences come from?' *International Journal of Game Theory* **42** (3): 613.

—— (forthcoming). 'Mentalism versus behaviourism in economics: a philosophy-of-science perspective.' *Economics and Philosophy*.

Dogramaci, S. (forthcoming). 'Knowing Our Degrees of Belief.' *Episteme*.

Domotor, Z. (1978). 'Axiomatizaton of Jeffrey Utilities.' *Synthese* **39**: 165-210.

Dreier, J. (1996). 'Rational preference: Decision theory as a theory of practical rationality.' *Theory and Decision* **40** (3): 249-76.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Center for the Study of Language and Information.

Dubois, D. and Prade, H. (1988). *Possibility Theory. An Approach to Computerized Processing of Uncertainty*. Plenum.

Easwaran, K. (2014). 'Decision Theory without Representation Theorems.' *Philosopher's Imprint* **14** (27): 1-30.

Eells, E. (1982). *Rational Decision and Causality*. Cambridge University Press.

Elliott, E. (forthcoming). 'Ramsey without Ethical Neutrality: A New Representation Theorem.' *Mind*.

Elliott, E., McQueen, K. and Weber, C. (2013). 'Epistemic Two-Dimensionalism and Arguments from Epistemic Misclassification.' *Australasian Journal of Philosophy* **91** (2): 375-89.

Ellsberg, D. (1961). 'Risk, ambiguity, and the Savage axioms.' *The Quarterly Journal of Economics* **75** (4): 643-69.

Eriksson, L. and Hájek, A. (2007). 'What are degrees of belief?' *Studia Logica* **86** (2): 183-213.

Fellner, W. (1961). 'Distortion of Subjective Probabilities as a Reaction to Uncertainty.' *Quarterly Journal of Economics* **75**: 670-89.

Fishburn, P. C. (1967). 'Preference-based definitions of subjective probability.' *The Annals of Mathematical Statistics* **38** (6): 1605-17.

—— (1970). *Utility theory for decision making*. John Wiley & Sons.

—— (1973). 'A Mixture-Set Axiomatization of Conditional Subjective Expected Utility.' *Econometrica* **41** (1): 1-25.

—— (1974). 'On the Foundations of Decision Making under Uncertainty'. In *Essays on Economic Behavior under Uncertainty*. M. S. Balch, D. L. McFadden and S. Y. Wu, Eds. North-Holland Publishing Company: 1-25.

—— (1975). 'A Theory of Subjective Expected Utility with Vague Preferences.' *Theory and Decision* **6**: 287-310.

—— (1981). 'Subjective expected utility: A review of normative theories.' *Theory and Decision* **13** (2): 139-99.

—— (1982). 'Nontransitive measurable utility.' *Journal of Mathematical Psychology* **26** (1): 31-67.

—— (1983). 'Transitive measurable utility.' *Journal of Economic Theory* **31** (2): 293-317.

—— (1994). 'Tales of a Radical Bayesian.' *Journal of Mathematical Psychology* **38** (1): 135-44.

Fishburn, P. C. and LaValle, I. H. (1988). 'Context-dependent choice with nonlinear and nontransitive preferences.' *Econometrica* **56** (5): 1221-39.

Fodor, J. (1975). *The Language of Thought*. Cromwell.

—— (1984). 'Semantics, Wisconsin Style.' *Synthese* **59**: 231-50.

—— (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.

Foley, R. (1992). 'The Epistemology of Belief and the Epistemology of Degrees of Belief.' *American Philosophical Quarterly* **29** (2): 111-21.

Fox, C. R., Rogers, B. A. and Tversky, A. (1996). 'Option Traders Exhibit Subadditive Decision Weights.' *Journal of Risk and Uncertainty* **13**: 5-17.

Fox, C. R. and Tversky, A. (1998). 'A Belief-Based Account of Decision Under Uncertainty.' *Management Science* **44** (7): 879-95.

Gaifman, H. and Liu, Y. (2015). *Decision Making without Miracles*. Unpublished manuscript. Retrieved from http://yliu.net/wp-content/uploads/state-dependent.pdf.

Ghirardato, P., Maccheroni, F., Marinacci, M. and Siniscalchi, M. (2003). 'A subjective spin on roulette wheels.' *Econometrica* **71**: 1897-908.

Gibbard, A. and Harper, W. L. (1978). 'Counterfactuals and Two Kinds of Expected Utility'. In *Foundations and Applications of Decision Theory*. C. A. Hooker, J. J. Leach and E. Francis, Eds. D. Reidel Publishing: 125-62.

Gilboa, I. (1987). 'Expected utility with purely subjective non-additive probabilities.' *Journal of Mathematical Economics* **16** (1): 65-88.

—— (1994). 'Can free choice be known?'. In *The Logic of Strategy*. C. Bicchieri, R. Jeffrey and B. Skyrms, Eds. Oxford University Press.

Gilboa, I. and Schmeidler, D. (1989). 'Maxmin Expected Utility with Non-unique Prior.' *Journal of Mathematical Economics* **18** (1989): 141-53.

Gonzalez, R. and Wu, G. (1999). 'On the Shape of the Probability Weighting Function.' *Cognitive Psychology* **38**: 129-66.

Goodman, N. D., Tenenbaum, J. B. and Gerstenberg, T. (2015). 'Concepts in a Probabilistic Language of Thought'. In *The Conceptual Mind: New Directions in the Study of Concepts*. E. Margolis and S. Laurence, Eds. MIT Press: 623-54.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T. and Danks, D. (2004). 'A Theory of Causal Learning in Children: Causal Maps and Bayes Nets.' *Psychological Review* **111** (1): 3-32.

Gul, F. and Pesendorfer, W. (2008). 'The Case for Mindless Economics'. In *The Foundations of Positive and Normative Economics*. A. Schotter, Ed. Oxford University Press: 3-39.

Hájek, A. (2003). 'What Conditional Probability Could Not Be.' *Synthese* **137** (3): 273-323.

—— (2008). 'Arguments for–or against–Probabilism?' *British Journal for the Philosophy of Science* **59** (4): 793-819.

Hájek, A. and Smithson, M. (2012). 'Rationality and indeterminate probabilities.' *Synthese* **187**: 33-48.

Halpern, J. Y. (2005). *Reasoning About Uncertainty*. MIT Press.

Hampton, J. (1994). 'The Failure of Expected-Utility THeory as a Theory of Reason.' *Economics and Philosophy* **10** (2): 195-242.

Harman, G. (1986). *Change in view*. MIT Press.

Harsanyi, J. (1977). 'On the rationale of the Bayesian approach: comments of Professor Watkins's paper'. In *Foundational Problems in the Special Sciences*. R. E. Butts and J. Hintikka, Eds. D. Reidel: 381-92.

Hausman, D. M. (2000). 'Revealed preference, belief, and game theory.' *Economics and Philosophy* **16** (01): 99-115.

Hawthorne, J. (2009). 'The Lockean Thesis and the Logic of Belief'. In *Degrees of Belief*. F. Huber and C. Schmidt-Petri, Eds. Springer: 49-74.

Hazen, G. B. (1987). 'Subjectively Weighted Linear Utility.' *Theory and Decision* **23**: 261-82.

Hedden, B. (2012). 'Options and the subjective *ought*.' *Philosophy Studies* **158**: 343-60.

Holton, H. (forthcoming). 'Intention as a Model for Belief'. In *Rational and Social Agency: Essays on the Philosopy of Michael Bratman*. M. Vargas and G. Yaffe, Eds. Oxford University Press.

Howson, C. and Urbach, P. (2005). *Scientific Reasoning: The Bayesian Approach*. Open Court Press.

Huber, F. (2009). 'Belief and degrees of belief'. In *Degrees of Belief*. F. Huber and C. Schmidt-Petri, Eds. Springer.

Huber, F. and Schmidt-Petri, C. (2009). *Degrees of belief*. Springer.

Jackson, F. (1998). *From metaphysics to ethics*.

—— (2009). 'Possibilities for representation and credence: two space-ism versus one space-ism'. In *Epistemic Modality*. A. Egan and B. Weatherson, Eds. Oxford University Press.

—— (2010). *Language, Names and Information*. Wiley-Blackwell.

Jeffrey, R. C. (1968). 'Probable knowledge.' *Studies in Logic and the Foundations of Mathematics* **51**: 166-90.

—— (1970). 'Dracula Meets Wolfman: Acceptance vs. Partial Belief'. In *Induction, Acceptance, and Rational Belief*. M. Swain, Ed. Reidel: 157-85.

—— (1974). 'Frameworks for Preference'. In *Essays on Economic Behavior under Uncertainty*. M. Balch, D. McFadden and S. Wu, Eds. North-Holland Publishing Company.

—— (1978). 'Axiomatizing the logic of decision'. In *Foundations and Applications of Decision Theory* Springer: 227-31.

—— (1983). 'Bayesianism with a human face.' *Testing Scientific Theories, Minnesota Studies in the Philosophy of Science* **10**: 133-56.

—— (1990). *The logic of decision*. University of Chicago Press.

Johanna, E., Jeleva, M. and Tallon, J.-M. (2012). 'Decision Theory under Ambiguity.' *Journal of Economic Surveys* **26**: 234-70.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge University Press.

—— (2002). 'Levi on causal decision theory and the possibility of predicting one's own actions.' *Philosophical Studies* **110**: 69-102.

Kahneman, D. and Tversky, A. (1979). 'Prospect theory: An analysis of decision under risk.' *Econometrica* **47** (2): 263-91.

Karni, E. (1993). 'Subjective Expected Utility Thoeory with State-Dependent Preferences.' *Journal of Economic Theory* **60**: 428-38.

Karni, E., Schmeidler, D. and Vind, K. (1983). 'On State Dependent Preferences and Subjective Probabilities.' *Econometrica* **51** (4): 1021-31.

Kearns, S. and Magidor, O. (2012). 'Semantic Sovereignty.' *Philosophy and Phenomenological Research* **85** (2): 322-50.

Kilka, M. and Weber, M. (2001). 'What Determines the Shape of the Probability Weighting Function under Uncertainty?' *Management Science* **47** (12): 1712-26.

Kochov, A. (2015). 'Time and No Lotteries: An Axiomatization of Maxmin Expected Utility.' *Econometrica* **83** (1): 239-62.

Koopman, B. O. (1940). 'The Bases of Probability.' *Bulletin of the American Mathematical Society* **46** (10): 763-74.

Krantz, D. H. and Luce, R. D. (1974). 'The interpretation of conditional expected-utility theories'. In *Essays on Economic Behavior under Uncertainty*. M. Balch, D. McFadden and S. Wu, Eds. North-Holland Publishing Company: 70-3.

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

Kreps, D. (1988). *Notes on the Theory of Choice*. Westview Press.

Kroon, F. (1987). 'Causal Descriptivism.' *Australasian Journal of Philosophy* **65**: 1-17.

Kyburg, H. E. (1992). 'Getting Fancy with Probability.' *Synthese* **90**: 189-203.

Levi, I. (1974). 'On Indeterminate Probabilities.' *The Journal of Philosophy* **71** (13): 391-418.

—— (1989). 'Rationality, Prediction, and Autonomous Choice.' *Canadian Journal of Philosophy* **19** (suppl.): 339-63.

—— (2000). 'Review of *The Foundations of Causal Decision Theory*.' *The Journal of Philosophy* **97** (7): 387-402.

—— (2007). 'Deliberation *Does* Crowd Out Prediction'. In *Homage à Wlodek; 60 Philosophical Papers Dedicated to Wlodek Rabinowicz*, edited by T. Rønnow-Rasmussen, B. Petersson, J. Josefsson and D. Egonsson. Lund University, Department of Philosophy. Retrieved from www.fil.lu.se/hommageawlodek.

Lewis, D. (1970). 'How to Define Theoretical Terms.' *The Journal of Philosophy* **67** (13): 427-46.

—— (1972). 'Psychophysical and theoretical identifications.' *Australasian Journal of Philosophy* **50** (3): 249-58.

—— (1974). 'Radical interpretation.' *Synthese* **27** (3): 331-44.

—— (1975). 'Languages and Language'. In *Minnesota Studies in the Philosophy of Science*. K. Gunderson, Ed. University of Minnesota Press: 3-35.

—— (1979). 'Attitudes De Dicto and De Se.' *The Philosophical Review* **88** (4): 513-43.

—— (1980a). 'Mad Pain and Martian Pain'. In *Readings in Philosophy of Psychology*. N. Block, Ed. Harvard University Press: 216-22.

—— (1980b). 'A Subjectivist's Guide to Objective Chance'. In *Studies in Inductive Logic and Probability*. R. C. Jeffrey, Ed. University of California Press: 263-93.

—— (1981). 'Causal Decision Theory.' *Australasian Journal of Philosophy* **59** (1): 5-30.

—— (1982). 'Logic for Equivocators.' *Nous* **16** (3): 431-41.

—— (1983). 'New work for a theory of universals.' *Australasian Journal of Philosophy* **61** (4): 343-77.

—— (1984). 'Putnam's Paradox.' *Australasian Journal of Philosophy* **62** (3): 221-36.

—— (1986). *On the Plurality of Worlds*. Cambridge University Press.

—— (1994). 'Reduction of Mind'. In *Companion to the Philosophy of Mind*. S. Guttenplan, Ed. Blackwell: 412-31.

Lichtenstein, S. and Slovic, P. (1971). 'Reversals of preference between bids and choices in gambling decisions.' *Journal of Experimental Psychology* **89** (1): 46.

—— (1973). 'Response-induced reversals of preference in gambling: An extended replication in Las Vegas.' *Journal of Experimental Psychology* **101** (1): 16.

Loar, B. (1981). *Mind and Meaning*. Cambridge University Press.

Luce, R. D. (1972). 'Conditional Expected, Extensive Utility.' *Theory and Decision* **3** (2): 101-6.

Luce, R. D. and Krantz, D. H. (1971). 'Conditional expected utility.' *Econometrica* **39** (2): 253-71.

Luce, R. D. and Suppes, P. (1965). 'Preference, Utility, and Subjective Probability'. In *Handbook of Mathematical Psychology*. D. R. Luce, R. R. Bush and E. H. Galanter, Eds. Wilely: 249-410.

Maccheroni, F., Marinacci, M. and Rustichini, A. (2006). 'Ambiguity Aversion, Robustness, and the Variational Representation of Preferences.' *Econometrica* **74** (6): 1447-98.

Machina, M. J. and Schmeidler, D. (1992). 'A More Robust Definition of Subjective Probability.' *Econometrica* **60** (4): 745-80.

Maher, P. (1993). *Betting on Theories*. Cambridge University Press.

—— (1997). 'Depragmatized Dutch Book Arguments.' *Philosophy of Science* **64** (2): 291-305.

Marcus, R. B. (1990). 'Some Revisionary Proposals about Belief and Believing.' *Philosophy and Phenomenological Research* **50**: 133-53.

Meacham, C. J. G. and Weisberg, J. (2011). 'Representation Theorems and the Foundations of Decision Theory.' *Australasian Journal of Philosophy* **89** (4): 641-63.

Millikan, R. G. (1989). 'Biosemantics.' *The Journal of Philosophy* **86** (6): 281-97.

—— (1990). 'Truth rules, hoverflies, and the Kripke-Wittgenstein paradox.' *The Philosophical Review* **9** (3): 323-53.

Narens, L. (1976). 'Utility-uncertainty trade-off structures.' *Journal of Mathematical Psychology* **13** (3): 296-322.

Neander, K. (2006). 'Content for Cognitive Science'. In *Teleosemantics*. G. McDonald and D. Papineau, Eds. Oxford University Press: 167-94.

Neilson, W. S. (2010). 'A simplified axiomatic approach to ambiguity aversion.' *Journal of Risk and Uncertainty* **41**: 113-24.

Nolan, D. (1997). 'Impossible Worlds: A Modest Approach.' *Notre Dame Journal of Formal Logic* **38**: 535-72.

Papineau, D. (1984). 'Representation and Explanation.' *Philosophy of Science* **51**: 550-72.

—— (1987). *Reality and Representation*. Basil Blackwell.

Pautz, A. (2013). 'Does Phenomenology Ground Mental Content?'. In *Phenomenal Intentionality*. U. Kriegel, Ed. Oxford: 194-234.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

—— (1990). 'Reasoning under uncertainty.' *Annual Review of Computer Science* **4** (1): 37-72.

Peterson, M. (2004). 'From Outcomes to Acts: A Non-Standard Axiomatization of the Expected Utility Principle.' *Journal of Philosophical Logic* **33**: 361-78.

Pettit, P. (1991). 'Decision theory and folk psychology'. In *Foundations of Decision Theory: Issues and Advances*. M. Bacharach and S. Hurley, Eds. Basil Blackwater: 147-75.

—— (1993). *The Common Mind*. Oxford University Press.

Pratt, J. W., Raiffa, H. and Schlaifer, R. (1965). *Introduction to Statistical Decision Theory*. McGraw-Hill.

Price, H. (2012). 'Causation, Chance, and the Rational Significance of Supernatural Evidence.' *The Philosophical Review* **121** (4): 483-538.

Putnam, H. (1980). 'Brains and Behavior'. In *Readings in Philosophy of Psychology*. N. Block, Ed. Harvard University Press: 24-36.

Qu, X. (2015). 'Purely subjective extended Bayesian models with Knightian unambiguity.' *Theory and Decision*: 1-25.

Quine, W. V. O. (1960). *Word and Object*. MIT Press.

Rabinowicz, W. (2002). 'Does Practical Deliberation Crowd Out Self-Prediction.' *Erkenntnis* **57**: 91-122.

Ramsey, F. P. (1927). 'Facts and Propositions'. In *The Foundations of Mathematics and other Logical Essays*. R. B. Braithwaite, Ed. Martino Publishing: 138-55.

—— (1931). 'Truth and probability'. In *The Foundations of Mathematics and Other Logical Essays*. R. B. Braithwaite, Ed. Routledge: 156-98.

Richter, M. K. (1975). 'Rational Choice and Polynomial Measurement Models.' *Journal of Mathematical Psychology* **12**: 99-113.

Ridge, M. (1998). 'Humean Intentions.' *American Philosophical Quarterly* **35**: 157-78.

Roberts, F. S. (1974). 'Laws of Exchange and Their Applications.' *SIAM Journal on Applied Mathematics* **26**: 260-84.

Samuelson, P. A. (1938). 'A note on the pure theory of consumer's behaviour.' *Economica* **5** (17): 61-71.

—— (1948). 'Consumption Theory in Terms of Revealed Preference.' *Economica* **15** (60): 243-53.

Sarin, R. and Wakker, P. (1992). 'A simple axiomatization of nonadditive expected utility.' *Econometrica* **60** (6): 1255-72.

Savage, L. J. (1954). *The Foundations of Statistics*. Dover.

Schervish, M. J., Seidenfeld, T. and Kadane, J. B. (1990). 'State-dependent utilities.' *Journal of the American Statistical Association* **85** (411): 840-7.

Schmeidler, D. (1989). 'Subjective probability and expected utility without additivity.' *Econometrica* **57** (3): 571-87.

Schmidt, U. (2002). 'Expected utility theory and alternative approaches'. In *Optimization and Operations Research*, edited by U. Derigs. University of Cologne.

Schneider, M. A. and Nunez, M. A. (2015). 'A simple mean-dispersion model of ambiguity attitudes.' *Journal of Mathematical Economics*.

Schwarz, W. (2014a). 'Against Magnetism.' *Australasian Journal of Philosophy* **92** (1): 17-36.

—— (2014b). *Decision theory for non-consequentialists*. Unpublished manuscript. Retrieved from http://www.umsu.de/papers/dt-for-noncons.pdf.

—— (2014c). *Options and Actions*. Unpublished manuscript. Retrieved from http://www.umsu.de/papers/options.pdf.

Schwitzgebel, E. (2002). 'A Phenomenal, Dispositional Account of Belief.' *Nous* **36** (2): 249-75.

—— (2013). 'A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box'. In *New Essays on Belief: Constitution, Content, and Structure*. N. Nottelmann, Ed. Palgrave Macmillan: 75-99.

Seidenfeld, T., Schervish, M. J. and Kadane, J. B. (1990). 'Decisions without Ordering'. In *Acting and Reflecting: The Interdisciplinary Turn in Philosophy*. W. Sieg, Ed. Kluwer Academic Publishers: 143-70.

—— (1995). 'A Representation of Partially Ordered Preferences.' *The Annals of Statistics* **23** (6): 2168-217.

Sen, A. (1973). 'Behaviour and the Concept of Preference.' *Economica* **40** (159): 241-59.

—— (1993). 'Internal Consistency of Choice.' *Econometrica* **61** (3): 495-521.

Seo, K. (2009). 'Ambiguity and Second-Order Belief.' *Econometrica* **77** (5): 1575-605.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

—— (1986). 'Savage revisited.' *Statistical Science* **1** (4): 463-501.

—— (2011). 'A betting interpretation for probabilities in Dempster-Shafer degrees of belief.' *International Journal of Approximate Reasoning* **52**: 127-36.

Shoemaker, S. (2003). *Identity, cause, and mind (expanded edition)*. Oxford University Press.

Soames, S. (1987). 'Direct Reference, Propositional Attitudes and Semantic Content.' *Philosophical Topics* **15**: 47-87.

Sobel, J. H. (1986). 'Notes on decision theory: Old wine in new bottles.' *Australasian Journal of Philosophy* **64** (4): 407-37.

—— (1997). 'Cyclical Preferences and World Bayesianisms.' *Philosophy of Science* **64** (1): 42-73.

—— (1998). 'Ramsey's Foundations Extended to Desirabilities.' *Theory and Decision* **44** (3): 231-78.

Spohn, W. (1977). 'Where Luce and Krantz do really generalize Savage's decision model.' *Erkenntnis* **11** (1): 113-34.

—— (1988). 'Ordinal Conditional Functions: A Dynamic Theory of Epistemic States'. In *Causation in Decision, Belief Change, and Statistics*. W. L. Harper and B. Skyrms, Eds. Kluwer: 105-34.

—— (1990). 'A General Non-Probabilistic Theory of Inductive Reasoning'. In *Uncertainty in Artificial Intelligence*. R. D. Shachter, T. S. Levitt, J. Lemmer and L. N. Kanal, Eds. North-Holland: 173-94.

Stalnaker, R. C. (1972). 'Letter to David Lewis'. In *Ifs*. W. L. Harper, G. A. Pearce and R. Stalnaker, Eds. Reidel: 151-2.

—— (1976). 'Propositions'. In *Issues in the Philosophy of Language: Proceedings of the 1972 Oberlin Colloquium in Philosophy*. A. F. MacKay and D. D. Merrill, Eds. Yale University Press: 79-91.

—— (1984). *Inquiry*. The MIT Press.

—— (1999a). 'Belief Attribution and Context'. In *Context and Content: Essays on Intentionality in Speech and Thought* Oxford University Press.

—— (1999b). 'Mental Content and Linguistic Form'. In *Context and Content: Essays on Intentionality in Speech and Thought* Oxford University Press: 225-41.

—— (2008). *Our Knowledge of the Internal World*. Oxford University Press.

Stampe, D. (1977). 'Toward a Causal Theory of Linguistic Representation'. In *Midwest Studies in Philosophy*. P. French, H. K. Wettstein and T. E. Uehling, Eds. University of Minnesota Press: 42-63.

Stigum, B. P. (1972). 'Finite state space and expected utility maximization.' *Econometrica* **40** (2): 253-9.

Stitch, S. (1992). 'What Is a Theory of Mental Representation?' *Mind* **101** (402): 243-61.

Suppes, P. (1969). 'The role of subjective probability and utility in decision-making'. In *Studies in the Methodology and Foundations of Science: Selected Papers from 1951 to 1969* Springer: 87-104.

—— (1974). 'The Measurement of Belief.' *Journal of the Royal Statistical Society Series B (Methodological)* **36** (2): 160-91.

—— (1994). 'Qualitative theory of subjective probability'. In *Subjective Probability*. G. Wright and P. Ayton, Eds. Wiley: 17-38.

Swoyer, C. (1991). 'Structural Representation and Surrogative Reasoning.' *Synthese* **87** (3): 449-508.

Tversky, A. (1975). 'A critique of expected utility theory: Descriptive and normative considerations.' *Erkenntnis* **9** (2): 163-73.

Tversky, A. and Fox, C. R. (1995). 'Weighing risk and uncertainty.' *Psychological Review* **102**: 269-83.

Tversky, A. and Kahneman, D. (1974). 'Judgment under Uncertainty: Heuristics and Biases.' *Science* **185** (4157): 1124-31.

—— (1981). 'The framing of decisions and the psychology of choice.' *Science* **211** (4481): 453-8.

—— (1992). 'Advances in prospect theory: Cumulative representation of uncertainty.' *Journal of Risk and Uncertainty* **5** (4): 297-323.

van Fraassen, B. (1990). 'Figures in a Probability Landscape'. In *Truth or Consequences*. J. Dunn and A. Gupta, Eds. Kluwer: 345-56.

—— (1995). 'Belief and the Problem of Ulyssess and the Sirens.' *Philosophical Studies* **77**: 7-37.

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

—— (1947). *Theory of Games and Economic Behavior, Volume 2*. Princeton University Press.

Wakker, P. P. (1989). 'Continuous Subjective Expected Utility with Non-Additive Probabilities.' *Journal of Mathematical Economics* **18**: 1-27.

—— (2004). 'On the composition of risk preference and belief.' *Psychological Review* **111**: 236-41.

—— (2010). *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press.

Wakker, P. P. and Tversky, A. (1993). 'An Axiomatization of Cumulative Prospect Theory.' *Journal of Risk and Uncertainty* **7** (7): 147-76.

Wakker, P. P. and Zank, H. (1999). 'State dependent expected utility for Savage's state space.' *Mathematics of Operations Research* **24** (1): 8-34.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall.

Weatherson, B. (2005). 'Can We Do Without Pragmatic Enchroachment?' *Philosophical Perspectives* **19**: 417-43.

—— (2012a). 'Knowledge, Bets and Interests'. In *Knowledge Ascriptions*. J. Brown and M. Gerken, Eds. Oxford University Press: 75-103.

—— (2012b). 'The Role of Naturalness in Lewis's Theory of Meaning.' *Journal for the History of Analytic Philosophy* **1** (10): 1-19.

—— (forthcoming). 'Games, Beliefs and Credences.' *Philosophy and Phenomenological Research*.

Weirich, P. (2004). *Realistic Decision Theory: Rules for nonideal agents in nonideal circumstances*. Oxford University Press.

—— (2015). *Models of Decision-Making*. Cambridge University Press.

Williams, J. R. G. (2007). 'Eligibility and Inscrutability.' *The Philosophical Review* **116** (3): 361-99.

—— (2008). 'The Price of Inscrutability.' *Nous* **42** (4): 600-41.

Williams, P. (1976). 'Indeterminate Probabilities'. In *Formal Methods in the Methodology of the Empirical Sciences*. M. Przelecki, K. Szaniawski and R. Wojcicki, Eds. Reidel: 229-46.

Williamson, J. (2010). *In Defence of Objective Bayesianism*. Oxford University Press.

Williamson, T. (2009). 'Reference, Inference and the Semantics of Pejoratives'. In *The Philosophy of David Kaplan*. J. Almog and P. Leonardi, Eds. Oxford University Press: 137-58.

Zynda, L. (2000). 'Representation Theorems and Realism About Degrees of Belief.' *Philosophy of Science* **67** (1): 45-69.