Decision Theory with a Human Face

Richard Bradley

August 2015

Contents

Preface

The aim of the book is to develop a decision theory that is tailored for 'real' agents; i.e. agents, like us, who are uncertain about a great many things and are limited in their capacity to represent, evaluate and deliberate, but which nonetheless want to get things right to the extent that they can. The book is motivated by two broad claims. The first is that Bayesian decision theory provides an account of the rationality requirements on 'unbounded' agents that is essentially correct and which is applicable in circumstances in which two main conditions are met: firstly, that agents are aware of all the options available to them and of all possibilities relevant to their evaluation and, secondly, that they are in a position to form precise judgements about the probability and desirability of these possibilities. The second is that there are many circumstances in which these conditions are not satisfied and hence in which the classical Bayesian theory is not applicable. A normative decision theory, adequate to such circumstances, would provide guidance on how 'bounded' agents should represent the uncertainty they face, how they should revise their opinions as a result of experience and how they should make decisions when lacking full awareness or precise opinions (that they have confidence in) on relevant contingencies. The book tries to provide such a theory.

So many people have helped me with this project over the many years it has taken to complete this project that I fear that I have will have forgotten many of them. It origins lie in my PhD dissertation done under the supervision of Richard Jeffrey, David Malament and Dan Garber. The influence of Dick Jeffrey on my thinking is hard to overestimate. The title of this book mirrors that of a paper of his—'Bayesianism with a human face'—in which he espoused the kind of heterodox Bayesianism that pervades my writing. To me he *was* the human face of Bayesianism.

Almost as much of an influence has been Jim Joyce, who I first met at a workshop on Jeffrey's work some twenty years ago. Big chunks of this book can be read as a dialogue with *The Foundations of Causal Decision Theory* and his subsequent work on Imprecise Bayesianism. Parts of it are based on ideas developed with coauthors on papers: in particular, Christian List and Franz Dietrich (chapter ?), Mareile Drechsler (chapter ?), Casey Helgeson and Brian Hill (chapter ?) and Orri Stefánsson (chapters ??). As with many others with whom I have worked over the years (including both PhD students and colleagues), I have largely lost my grip on which ideas are mine and which are theirs (if such a separation can meaningfully be made). It is an unfortunate irony that the ideas that you most thoroughly absorb are often the ones whose origins you forget.

I have been at the LSE for most of my career and it has provided the best

possible intellectual environment for writing the book. The weekly seminars of LSE Choice Group have provided an invaluable forum for presenting ideas and acquiring new ones and its member a source of support, both intellectual and more generally. Outside the LSE, Philippe Mongin and John Broome were both very supportive at crucial moments.

A number of people read parts of the book manuscript including Magda Osman, Seamus Bradley, Conrad Heilmann, Hykel Hosni, Susanne Burri, Philippe van Baeyshuis and Silvia Milano. Orris Stefánsson not only read an entire draft, but has been a wonderful interlocutor on its contents over many years. Katie Steele, Anna Mahtani, Jim Joyce, Wlodek Rabinowicz and Christian List provided valuable feedback on individual chapters at a workshop organised by Christian.

I am grateful the AHRC for their support both in the form of a grant (AH/I003118/1) to work on the book and for a grant for a project on *Managing Severe Uncertainty* (AH/J006033/1) the fruits of which are contained in the last part of the book.

Finally, I am deeply grateful to my family, and especially my wife Shura, for putting up with me over the last few years. There have been a good number of 'holidays' and weekends lost to book writing, not to mention grumpiness when nothing seemed to progress, but their patience and support has been undiminished by it all.

Introduction

Decision problems abound. Consumers have to decide what products to buy, doctors what treatments to prescribe, hiring committees what candidates to appoint, juries whether to convict or acquit a defendant, aid organisations what projects to fund, monetary policy committees what interest rates to set, and legislatures what laws to make. The problem for descriptive decision theory is to explain how such decisions are made. The problem that normative decision theory faces, on the other hand, is how individuals and groups facing choices such as these should make their decisions: how should they evaluate the alternatives before them, what criteria should they employ, what procedures should they follow?

As these examples illustrate decisions have to be made in a wide variety of contexts and by different kinds of decision makers. A pervasive feature however is the uncertainty that the decision maker faces: uncertainty about the state of the world, about the alternatives available to them, about the possible consequences of making one choice rather than another and indeed about how to evaluate these consequences. Dealing with this uncertainty is perhaps the most fundamental challenge we face in making a decision.

In the last 100 years or so an impressive body of work on this issue has emerged. At its core stands Bayesian decision theory, a mathematical and philosophical theory of both rational belief and change of belief and of rational decision making under uncertainty. The influence of Bayesian thinking pervades this book, to which the amount of space devoted to examining and criticising it will attest. Indeed I regard Bayesian decision theory as essentially the correct theory for a type of decision situation, namely those in which the decision maker is in what might be called a state of opinionated equilibrium. To be in such a state, the decision maker must have reached a judgement about all matters of relevance to the decision at hand and have completely thought through the implications of these judgements, so that as a set they are consistent and complete. Unfortunately, we rarely find ourselves in a state of this kind. So although Bayesian decision theory is a very good starting point for thinking about rational decision making, it's not the complete story.

This is for two main reasons. Firstly, Bayesian theory assumes that decision makers are 'unbounded': rational, logically omniscient and maximally opinionated. Rational in that their attitudes - beliefs, desires and preferences - are consistent both in themselves and with respect to one another; logically omniscient because they believe all logical truths and endorse all the logical consequences of their attitudes; and opinionated because they have determinate beliefs, desires and preferences regarding all relevant prospects. All of these assumptions can be criticised on grounds of being unrealistic: human decision

makers, for instance, are unlikely to satisfy them for anything but a very small sets of prospects. Some of them can also be criticised on normative grounds. It is surely not required of us, for instance, that we have opinions about everything, nor that we are aware of all possibilities.

Secondly, by formulating the notion of a decision problem in a particular way, Bayesian decision theory excludes many of the kinds of uncertainty mentioned before. Indeed it essentially restricts uncertainty to our knowledge of the state of the world, leaving out the uncertainty we face in judging how valuable consequences of actions are and the uncertainty we face as to the effect of our interventions in the world. Furthermore it assumes that all uncertainty is of the same kind or severity, one that can be captured by a (single) probability measure on the possible states of the world. But we are often so unsure of things that we cannot even assign probabilities to them. It follows that Bayesianism is incomplete as a theory of rationality under uncertainty.

The main aim of the book, as its title suggests, is to develop a decision theory that is tailored for 'real' agents facing uncertainty that come in many forms and degrees of severity. By real agents I mean those who, like us, have limited skills and restricted time and other computational resources that make it impossible and undesirable that they should form attitudes to all contingencies or to think through all the logical consequences of what they believe, but which nonetheless get things right to the best of their ability and who employ quite sophisticated reasoning to this end. Humans, for instance, are not just capable of representing their environment as they find it, but also of reflecting prospectively about how it might and would be if certain contingencies turned out to be true or if they were to perform particular actions. And of reflecting retrospectively on experience and in particular on the outcomes of past actions, enabling them to improve their understanding of the world and the effect of their interventions in it. An examination of these abilities takes us into areas neglected by Bayesianism, such as the study of hypothetical reasoning and of reasoned preference change.

The desirability of moving in the direction of greater realism is, not surprisingly, widely recognised. But the way in which I want to do so is different from the direction taken in, for instance, behavioural economics and the psychology of judgement and choice. For the aim is not to describe the way in which we do in fact evaluate prospects and make decisions, but to prescribe how we should, given our limitations and constraints. *The project is thus of giving not a descriptive theory of bounded rationality, but a normative theory of rationality for the bounded.*[1]

A decision theory that aims to play this kind of normative role must address itself to the sorts of agents that we are and the sorts of decision problems we face, take as its starting point the resources and judgements that are available to them to deal with it. This means firstly that decision makers should be free to represent the decision problem to themselves as they see fit, in the light of the resources that they can bring to bear on the problem. And secondly that guidance should be provided on forming and revising judgements, as well as on making decisions; guidance that is appropriate to kind of uncertainty they face. It does so by doing what philosophers do best: proposing and examining

[1] These projects overlap to some degree of course. Indeed Herbert Simon's canonical work (see Simon (1957, 1986, 1990)) addresses both normative and desriptive issues, as does the more recent work of Paul Weirich (2004). This book is complementary to their work, but its foucs us much further away from the details of cognitive mechanisms.

candidate principles of rational belief, desire and choice that bounded agents can use to bring order to their deliberations, both prospective and retrospective, and to their actions.

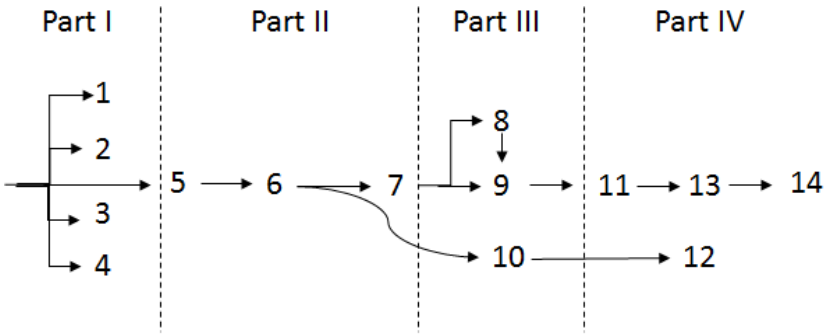
It's an enterprise that is at once very ambitious and quite modest. Ambitious because it aims at finding rationality principles of very general scope, applicable to the deliberations of many different kinds of decision situations and decision makers. Modest, because these principles impose only conditions of consistency on agents. The theory does not attempt to dictate whether we should believe, value or do any specific thing, but only which patterns of believing, valuing and doing are permitted. Rationality alone cannot decide for us what to think or to do, but it can support us in our attempts to do so. In pursuing this project my perspective is always that of the decision maker. That is, I take the task to be of prescribing what the agent should do, given their understanding of the decision problem, their preferences and their beliefs. It is not to tell the agent what the decision problem really is or what preferences or beliefs they should have beyond considerations of consistency. These are perfectly good questions to address, but they are not the ones that will occupy me. The subjectivism inherent in this approach will make some uneasy, but I will do my best to vindicate it as we go along.

Book Outline

The book is divided into four parts. The first part introduces the basics of Bayesian theory and then looks a range of philosophical questions about its foundations, interpretation and application, including the framing of decision problems (ch. 1), the nature of rationality and the interpretation of probability and utility (ch. 2), and the classification of forms of uncertainty (ch. 3). It also assesses the role of representation theorems in motivating decision theories, looking in detail at Savage's version by way of illustration (ch. 4).

Part II of the book is devoted to developing a theory of prospectively rational agency, the kind of rationality characteristic of agents that not only represent and evaluate the current state of their environment, but also the state that might obtain or would obtain if they (or others) were to intervene in it in some way. The basic building blocks for the account are provided by Richard Jeffrey's version of Bayesian decision theory and the representation theorems for it due to Ethan Bolker and James Joyce, but the theory is extended to the treatment of conditional attitudes (ch. 6) and then to conditionals (ch.7) by enriching the set of prospects and proposing rationality conditions on belief and desire appropriate to them. Each set of claims is supported by a representation theorem showing how the quantitative claims under consideration have foundations in rationality constraints on relational attitudes of belief and desire.

Part III considers how a prospectively rational agent interacts with the world, applying the framework developed in Part II. There are three aspects to this. The first is the semantic issue of how the agent represents the prospects that are the objects of her attitudes. The second is the issue of how agents should evaluate their own interventions in the world and make decisions on the basis of such evaluations. The third is the effect of experience on the agent's attitudes, i.e. of how she learns from experience. Ch. 9 deals with the first by explaining how conditional prospects are modelled in multidimensional possible world semantics and showing that this way allows for non-trivial satisfaction of



the rationality claims made in the second part of the book. Ch. 10 gives the core account of decision making under risk and uncertainty, showing that the theories of von Neumann and Savage can be derived within the framework of the book in the presence of the special assumptions about the objects of choice and deriving a particular formulation of causal decision theory. Ch. 11 develops the Bayesian theory of learning, defending forms of conditionalisation appropriate to a variety of different learning experiences.

Part IV develops an account of the rationality of bounded agents: agents that lack full awareness and who are not maximally opinionated. Ch. 12 defends a version of Imprecise Bayesian, seeking foundations for it in the notion of coherently extendable preferences and showing how it can deal with unawareness. But it also raises a number of challenges for Imprecise Bayesianism; challenges that are taken up in subsequent chapters. Ch. 13 examines how an agent with imprecise beliefs and desires changes her mind in response to experience, developing a broadly Bayesian account of attitude formation and withdrawal that complements the standard accounts of attitude change. Ch. 14 looks at how such an agent might make decisions, comparing the strategy of making up one's mind with that of applying an alternative decision rule. Special attention is paid to considerations of caution and to the question of whether and how caution can be rationalised. The final chapter argues that considerations of confidence must be drawn on to handle the challenges to Imprecise Bayesianism, to provide a basis for both belief revision and decision making that is appropriately sensitive to the agent's state of uncertainty both about what to believe and what to desire.

Readers can make their way through the book in different ways, depending on background and interests. Those impatient with philosophical preliminaries and a good background in decision theory can jump right to part II. Part II is fairly self-contained but parts III and IV depend on it. With part II under your belt it suffices to read chapter 9 in order to read part IV. Figure summarises the dependencies between chapters.

Some parts of the book are more technical than others. When these technical details are essential to the argument, then I try to explain them fully. When they are not, I have placed them in a starred section.

Part I

Rationality, Uncertainty and Choice

Chapter 1

Framing Decision Problems

Decision theory begins with decision problems. Decision problems arise for *agents*: entities with the resources to represent, evaluate and change the world around them in various different ways, typically within the context of ongoing personal and institutional projects, activities or responsibilities. These projects together with the environment, both natural and social, provide the givens for the decision problems agents face: their material resources for acting, their information and often their standards for evaluating outcomes, as well as the source of the problems they must respond to. Social scientists hold very different views about the relative importance of the different aspects of this background and the decisions that are made within it, but few would doubt that choices made by consumers, doctors, policy makers and so on have the power to shape the course of events.

To face a genuine decision problem, an agent must have options: actions that they are capable of performing and, equally, of foregoing if they so choose. To get an idea of what sorts of things count as a decision problem, let's look at a few examples.

1. *Take a bus?* You have an appointment that you don't want to miss. If you walk you will arrive late. If you take the bus and the traffic is light, you should arrive ahead of time. On the other hand if the traffic is heavy then you will arrive very late, perhaps so late that the appointment will be lost. Is it worth risking being late?

2. *Health Insurance?* You are presently in good health but know that if you were to fall ill you might not be able to continue to earn an income and, in the worst case, you might not be able to afford the health care you need. By buying health insurance you can ensure that you have all the care you need. But its expensive and if your health remains good, the money is wasted. Is it worth insuring yourself?

3. *Free condoms.* By supplying condoms free, rates of transmission of venereal disease can be considerably reduced. But there is the possibility that it will also encourage sexual activity thereby partially, or perhaps even completely, offsetting the benefits of a decreased transmission rate by virtue of the increase in the number of sexual liaisons. Should they be supplied free of charge?

	Heavy traffic	*Light traffic*
Take a bus	Arrive late Pay for a ticket	Arrive early Pay for a ticket
Walk	Arrive a little late No ticket needed	Arrive a little late No ticket needed

Table 1.1: Take a Bus?

	States			
Options	State S_1	State S_2	...	State S_n
α	A_1	A_2	...	A_n
β	B_1	B_2	...	B_n
...	...	...	...	...
γ	C_1	C_2	...	C_n

Table 1.2: State-Consequence Matrix

4. *Vaccinations.* A vaccine has been developed for cervical cancer, a fairly common type of cancer with a high mortality rate. The vaccine is expensive, but if it is developed as part of a large-scale vaccination programme, the costs are not exorbitant. The vaccine does however have severe side-effects in very rare cases (less than 1 in 100,000). Should the government offer the vaccine to everyone, actively encourage them to be vaccinated or even introduce compulsory vaccination?

Decision problems like these can be described in the following way. A decision maker has one or more options before them. The exercise of each option is associated with a number of possible consequences, some of which are desirable from the perspective of the decision maker's goals, others are not. Which consequence will result from the exercise of an option depends on the prevailing features of the environment: whether traffic is light or heavy, how much it is raining, whether you fall ill, and so on.

Let us call the set of environmental features relevant to the determination of the consequences of the exercise of any of the options, a state of the world. Then a decision problem can be represented by a matrix showing, for each available option, the consequence that follows from its exercise in each relevant state of the world. In our first example, for instance, taking the bus has the consequence of having to buy a ticket and arriving late in the event of heavy traffic and paying for a ticket and arriving early in the event of light traffic. This decision problem can be represented in a simple way as in Table 1.1 where the consequences of the available options listed in the rows of the table, for each of the states listed in the columns, are given in the table cells.

More generally suppose that α, β, ..., and γ are the options open to the decision maker and that S_1 through S_n are n possible states of the world (these must be mutually exclusive and exhaust all the possibilities). For any option γ, let C_1 through C_n be the n consequences that might follow from exercising it. Then a decision problem can be represented by a state-consequence matrix of the kind displayed in Table 1.2:

	Probabilities of States		
Options	$P(S_1)$	...	$P(S_n)$
α	$U(A_1)$	...	$U(A_n)$
β	$U(B_1)$	...	$U(B_n)$
...	...	...	...
γ	$U(C_1)$	...	$U(C_n)$

Table 1.3: Probability-Utility Matrix

Given a decision problem of this kind, standard decision theory says that the decision maker should choose the option whose exercise has the *greatest expected benefit*, where benefit is relative to the decision maker's evaluation of the desirability of the possible consequences of her actions. If she knows what the actual state of the world is then she should simply pick the option with the most desirable consequence in that state. Typically however the decision maker will be uncertain as to which is the actual state. In this case, she must consider how probable it is that each of the states is the case and pick the option whose expected desirability is greatest given these probability judgements.

For instance suppose that I consider the probability of heavy traffic to be one-half and the benefit or utility of the various possible consequences to be as below:

	0.5	0.5
Take a bus	-2	1
Walk	-1	-1

Then the expected benefit of taking the bus is a probability weighted average of the benefits of its possible consequences, i.e. $(-2 \times 0.5) + (1 \times 0.5) = -0.5$. On the other hand, walking has a certain benefit of -1. So in this case I should take the bus. But had the probability of heavy traffic been a lot greater, walking would have been the better option.

The next couple of chapters will be devoted to qualifying, expanding and commenting on the claim illustrated in this simple example, namely that we should pick the option that maximises expected utility. But before we do so, it will be helpful to express it more formally so that the core content is clear. Let P be a probability measure on the states of the world and U a utility measure on consequences (we will say more about what these measures are and where they come from in due course). Then a state-consequence matrix, such as that of Table 1.2, induces another matrix in which options appear as random variables: functions that assign a utility value to each state of the world (intuitively, the utility of the consequence of exercising the option in question in that state). This matrix is given in Table 1.3. So represented, each option has an expected value that is jointly determined by the functions U and P. For instance, the expected value of option γ, denoted by $\mathbb{E}(\gamma)$, is given by $U(C_1) \cdot P(S_1) + ... + U(C_n) \cdot P(S_n)$.

More generally, if the number of possible states of the world is finite:[1]

$$\mathbb{E}(\gamma) = \sum_{i=1}^{n} U(C_i).P(S_i)$$

Now what decision theory says is that rational agents should choose the option with the highest expected value. This is known as the **maximisation of expected utility hypothesis**.

The maximisation hypothesis forms the core of Bayesian decision theory, together with claims about how uncertainty should be represented and resolved through learning (respectively discussed in chapters 3 and 10). I will argue that this hypothesis is essentially correct for cases in which we can adequately represent the decision problem we face in a manner similar to that of Table 1.3, i.e. when we can display the problem in a state-consequence matrix and can reach probability and utility judgements on all the relevant factors displayed in it. When we cannot (which is quite often the case) then the theory is not false, but inapplicable, and much of the last part of this book will be devoted to answering the question as to what we do then. But for the moment our focus will be on understanding what the maximisation of expected utility hypothesis says, examining in this chapter how decision problems should framed and, in the next, how the hypothesis should be interpreted and what notion of rationality it presupposes.

1.1 Framing Decisions

Decision theory makes a claim about what option(s) it is rational to choose when the decision problem faced by the agent can be represented by a state-consequence matrix of the kind exemplified by Table 1. It is very important to stress that the theory does not say that you *must* frame decision problems in this way. Nor does it say that agents *will* always do so. It just says that *if* they are framed in this way, then only options which maximise expected benefit should be chosen. Nothing precludes the possibility that the same decision situation can or must be framed in different ways. This is true in more than one sense.

Firstly, it may be that the problem is not naturally represented by a state-consequence decision matrix. As John Broome Broome (1991) points out, the consequences of an action may distributed across different times or places or people, as well as across states. The desirability of ordering a cold beer or not, for instance, will depends on the location of its consequences: it's good if the beer is served to me, in the evening, with a smile and when I have not had a few too many already; bad when it's for my children, or first thing in the morning, or during a philosophy lecture. In this case my decision problem is better represented by a matrix that associates each action and relevant combination of locations (person, time, place, etc.) with a consequence, rather than a simple state-consequence matrix.

Secondly, the problem may not be representable by any kind of decision matrix at all because we are unable to identify the various elements of it: what our options are, what the relevant factors are that determine the consequence

[1] The restriction to a finite number of states of the world is made for simplicity, but the expected value will still be well defined even if we drop it.

of each option, or what the consequences are of exercising one or another of the identified options when these factors are present. In particular we may not be able to assign a determinate consequence to each state of the world for each option if the world is non-deterministic or if we cannot enumerate all the relevant conditions, a problem that I will term option uncertainty. This problem is discussed in detail in section 3.2.

Finally, it is typically possible to represent the decision problem one faces by any number of different decision matrices that differ in terms of the features of the problem that they explicitly pick out. This is true even if we just confine attention to state-consequences matrices (as I shall do), for our state of uncertainty can be more or less elaborately described by representing more or fewer of the contingencies upon which our decision might depend.

This last point raises the question of whether all such representations are equally good, or whether some are better (or worse) than others. There are two claims that I want to make in this regard: firstly that not all representations of a decision problem are equally good and, secondly, that many representations are nonetheless permissible. This latter point is of some importance because it follows that an adequate decision theory must be 'tolerant' to some degree of the manner in which a problem is represented and that the solution it gives to a decision problem should be independent of the choice of representation

Let us start with the first claim, that some representations of a problem are better than others. A representation of a decision problem should help us arrive at a decision by highlighting certain features of the problem and in particular those upon which the decision depends. What makes one way of framing the problem better than another is simply that it is more helpful in this regard. There are at least two considerations that need to be traded off when talking about the usefulness of a representation: the quality of the decisions likely to be obtained and the efficiency of obtaining them. Let me say something about them both.

Quality: To make a good decision, a decision maker must give appropriate weight to the factors upon which the decision depends. In deciding whether to take an umbrella or not, for instance, I need to identify both the features of the possible outcomes of doing so that matter to me (e.g. getting wet versus staying dry) and the features of the environment upon which these outcomes depend (e.g. the eventuality of rain). Furthermore I need to determine how significant these features are: how desirable staying dry is relative to getting wet, how probable it is that it will rain, and so on. If my representation of the decision problem is too sparse, I risk omitting features that are relevant to the decision. If I omit possible weather states from my representation of the umbrella-taking decision, for instance, then I may fail to take into account factors (in particular the probability of rain) upon which the correctness of the decision depends. So, *ceteris paribus*, a representation that includes more relevant features will be better than one that does not.

Efficiency: One way of ensuring that no relevant features are omitted is simply to list *all* the features of possible outcomes and states of the world. But drawing up and making use of such a list is clearly beyond our human capabilities and those of any real agents. Reaching judgements costs in terms of time

and effort. If we try to consider all possible features of the world we will simply run out of time and energy before making a decision. A framing that delivers accuracy but is so complex that it is impossible to specify all the required inputs and to compute the expected utilities is clearly not of much use. More generally, representations that include too many features will result in inefficient decision making requiring more resources than is justified (what level of resources is justified will of course depend on what is at stake). So, *ceteris paribus*, a simpler representation will be better than a more complicated one.

Achieving a good trade-off between quality and efficiency is not just a matter of getting the level of complexity right. It is also a matter of identifying the most useful features to represent explicitly. It is useful to represent a feature if it is (sufficiently) relevant to the decision and if we can determine what significance to attach to it. A feature of the state of the world or of a consequence is relevant to a decision problem if the choice of action is sensitive to values that we might reasonably assign to this feature (its probability or utility). More precisely, one feature is more relevant than another just in case the expected values of the various actions or options under consideration are more sensitive to changes in the values of the former than the latter. For instance, whether it is desirable to take an umbrella with me or not will be sensitive to the probability of rain, but not sensitive at all to the probability of a dust storm on Mars. Likewise it is sensitive to the utility of my getting wet but not to my getting hungry, since my getting wet depends causally on the taking of the umbrella but not my getting hungry. So a good representation of my decision problem will include weather states and 'wet-dry' consequences, but not Martian dust storm states or 'hungry' consequences.

The second aspect of usefulness is equally important. A representation should be appropriate to our informational resources and our cognitive capabilities in specifying features of the environment that we are capable of tracking and features of consequences that we are capable of evaluating. If the weather is relevant to my decision as to whether to take an umbrella or not, but I am incapable of reaching a judgement as to whether it is likely to rain or not (perhaps I have no information relevant to the question or I don't understand the information I have been given) then there is little point in framing the decision problem in terms of weather contingencies. A good representation of a problem helps us to bring the judgements we are able to make to bear on the decision problem.

It follows that whether a framing is a useful one or not will depend on properties of the decision maker (and in more than one way). Firstly whether the features of the problem it represents are relevant depends on what matters to the decision maker and hence what sort of considerations her decisions will be sensitive to. And secondly whether a representation facilitates decision making will depend on the cognitive abilities and resources of the decision maker. Both of these will vary from decision maker to decision maker and from one time and context to another.

It is clearly desirable therefore that a decision theory be 'representation tolerant' to as great a degree as possible, in the sense of being applicable to a decision problem irrespective of how it turns out to be useful for the decision maker to represent it. Not all decision theories are equal in this regard. On the contrary, as we shall see in the next section, some impose quite severe restrictions on how a decision problem must be represented if the theory is to be used and

hence make considerable demands on the decision maker in terms of the number and complexity of judgements that they must reach. Given our aim of a decision theory with a human face, this feature will count heavily against such theories.

1.2 Savage's Theory

The modern theory of decision making under uncertainty has its roots in 18th century debates over the value of gambles, with Daniel Bernouilli (in **?**) giving the earliest precise statement of something akin to the principle of maximising expected utility. The first axiomatic derivation of an expected utility representation of preferences is due to Frank Ramsey Ramsey (1990) whose treatment in many way surpasses those of later authors. But modern decision theory descends from Savage, not Ramsey, and it is in his book *The Foundations of Statistics* that we find the first simultaneous derivation of subjective probabilities and utilities from what are clearly candidate rationality conditions on preference.

It is to Savage too that we owe the representation of decision problems faced by agents under conditions of uncertainty that was described at the beginning of the chapter and that is now standard in decision theory. Its cornerstone is a tripartite distinction between states, consequences and actions. Consequences are the features of the world that the agent cares about and seeks to bring about or avoid by acting; they are, he says, "anything that may happen to an agent" or "anything at all about which the person could possibly be concerned" (**?**, p. 13-14). States are those features of the world that are outside of the agent's control but determine what consequence follows from the choice of action. Actions are the link between the two; formally, for Savage, they are just functions from states to consequences.

Although the tripartite distinction between states, consequences and actions is natural and useful, Savage's theory imposes some quite stringent conditions on how these objects are to be conceived. Firstly, in order that decision problems be representable by state-consequence matrices of the kind given in Table 1.2, he requires that the states of the world suffice to determine the consequence of a choice of action (I will discuss this in more detail in the next chapter). Secondly, he requires that the states themselves be causally and probabilistically independent of the action performed. And thirdly, he requires that the desirability of consequences be independent both of the state of the world in which they are realised and of the action. Jointly these assumptions imply that actions differ in value only insofar as they determine different ordered sets of consequences.

To ensure that the second two conditions hold, Savage suggested that consequences be maximally specific with regard to all that matters to the agent so that there be no uncertainty about how beneficial or desirable the consequence is that derives from uncertainty about the state of the world. It follows that the states themselves must be maximally specific, "leaving no relevant aspect undescribed" (**?**, p. 14), for if this were not the case then there could be features of the consequences of actions that matter to the agent but which are not determined by the prevailing state. It follows that when an agent regards a Savage-style action as open to them, they must take it that they can bring it about that a maximally specific consequence will obtain conditional on each maximally specific state of the world prevailing. This is rather different to what we colloquially understand by an action. When I must choose between walking

	Good health	Poor health
Purchase Insurance	Earn full annual income Make policy payments	Reduced income Insurance pays out
Don't Purchase	Earn full annual income No policy payments	Reduced income No payout

Table 1.4: Insurance Purchase

to the shops or taking the bus, as in the decision problem represented by Table 1.1, I do not do so in the light of anything like full knowledge of the consequences, in each possible state of the world, of these actions. My understanding of them is inevitably coarse-grained to some extent. It would seem then that Savage's theory is not well-suited to agents like us, who cannot typically represent decision problems in the way required for application of his theory.

Savage was perfectly aware of this objection and drew an important distinction between small-world and grand-world decision problems. Grand-world decision problems are ones which have consequences that are maximally specific with regard to all matters of concern to the agent; small-world problems are ones with coarse grained specifications of states and consequences. Although his theory is designed for grand-world problems, Savage argued that it could nonetheless be applied to small-world problems, so long as we ensure that the coarse-grained representation of the decision problem had sufficiently similar properties to the fine-grained one that it could be given a numerical representation by a probability-utility matrix of the kind exhibited in Table 1.3. For this, the two conditions of probabilistic independence of states from actions and desirabilistic independence of consequences from states are essential.

The upshot is that Savage's theory is far from being representation tolerant in the way that I argued was desirable. Indeed it is quite easy to fall foul of his constraints. Suppose that we are deciding whether to purchase health insurance for the coming years and that we represent our decision problem by the state-consequence matrix displayed in Table 1.4. So represented it looks like a purely financial decision, that can be made on the basis of the expected incomes associated with the two possible acts. It is quite conceivable, however, that the value we attach to income depends on our state of health. We might need more money if our health is poor, for instance, in order to buy services that we can no longer provide for ourselves. This would be a reason to value a particular income more highly if it is gained under poor health than if it is gained under good health. Alternatively, we may get less enjoyment from money when our health is poor, and so we would value it less. Either way, in order to use Savage's theory to make a decision as to whether to purchase insurance or not, in a way that appropriately reflects the sensitivity of the desirability of money on health states, the decision problem must be reframed.

The obvious way of doing this is to take the consequences of options to be combinations of outcomes and the states in which they are realised. The act of buying health insurance, for instance, may be said to have the consequence "Earn full income, make policy payments, enjoy good health" in good-health states and the consequence "Earn reduced income, make policy payments, enjoy poor health" in ill-health ones. However, for reasons that we will examine

more closely later on, Savage requires that consequences and states be logically independent. So he is forced to insist that decision makers describe the consequences of their actions in terms which eliminate the sensitivity of their value to the state of the world. This is not straightforward. The way in which income varies with health-states is likely to be mediated by an enormous number of variables, including the amount of support that can be expected from friends and family, the services provided by the state, charities or other institutions to help those in poor health, and one's level of psychological wellbeing. All of these would have to be specified in order for the act of purchasing health insurance to have a state-independent consequence in each health state. We are rarely able to do this.

A second problem in our example concerns the description of the relevant states, for the purchase of health insurance can have a causal effect on how much care one takes of one's health, so that the probability of good health is not independent of the purchase of health insurance. This problem of moral hazard, as it called, plagues insurance markets. When you sell someone fire insurance, for example, you change their incentives in such a way as to make it more probable that a fire will occur. Knowing that they will be reimbursed if a fire occurs, individuals may be less careful. In extreme cases, when the value of the policy is high enough, they may even commit arson. Insurance companies have to be very careful when selling fire insurance not to underestimate their exposure. To eliminate the possibility of moral hazard the decision problem has to be reframed. In our example this would require identifying all those factors (genetic, environmental, historical) mediating the relationship between purchases of health insurance and health states, combinations of which would serve as states in the reframed decision problem. This can be very difficult to do.

1.3 Bolker-Jeffrey Decision Theory

We have seen that apply Savage's theory, three conditions must be satisfied: acts must determine definite consequence in each state of the world; the desirability of each of the consequences must be independent both of the state of the world in which it obtains and the action that brings it about; and the states of world must be probabilistically independent of the choice of action. Taken separately it is often possible to ensure that for all practical purposes these conditions are met by being careful about how the decision problem is framed. But ensuring that all three are satisfied at the same time is very difficult indeed since the demands they impose on the description of the decision problem pull in different directions. For instance, the first condition is most easily met by coarsening the description of outcomes, but the second requires refining them.

This problem provides strong grounds for turning our attention to a rival version of Bayesian decision theory that is due to Richard Jeffrey and Ethan Bolker. Jeffrey (1990/1983) makes two modifications to the Savage framework. First, he recognises that the distinction between states and consequences is both context and agent dependent: that it will rain is a possible state of interest to a farmer, but a consequence for a shaman with a rain dance repertoire; that there will be flooding in low-lying areas is a possible state of the world from the perspective of a person buying a house, but a consequence from the point

	States		
Options	S_1	...	S_n
α	$P_\alpha(S_1), U(\alpha, S_1)$	...	$P_\alpha(S_n), U(\alpha, S_n)$
...	...	...	...
γ	$P_\gamma(S_1), U(\gamma, S_1)$	...	$P_\gamma(S_n), U(\gamma, S_n)$

Table 1.5: Act-Dependent Consequence Matrix

of view of environmental policy. So instead of distinguishing between objects of belief (states) and those of desire (consequences), he takes the contents of all of the decision maker's attitudes to be propositions. States and consequences become propositions that can be distinguished pragmatically, but not logically. This small modification has a very important implication. Since events and consequences are logically interrelated in virtue of being the same kind of object, the desirabilities of consequences are necessarily state-*dependent* in Jeffrey's framework. This means that his theory is not subject to the second of the restrictions required for Savage's.

The second modification that Jeffrey makes is more contentious and requires a bit of explanation. If he followed Savage in defining actions as arbitrary functions from partitions of events to consequences, the fact that in principle any proposition could serve as a consequence would imply an explosion in the size of the set of actions. But Jeffrey argues that many of the actions so defined would be inconsistent with the causal beliefs of the decision maker.[2] Someone may think they have the option of making it true that if the traffic is light they will arrive on time for their appointment, and if it's heavy they will arrive late, but not believe that it is possible to make it true that if the traffic is light they arrive late, and if it's heavy they arrive on time. So instead Jeffrey conceives of actions as simply those propositions that can be made true at will, characterised for decision purposes by the probabilities and utilities (or desirabilities as Jeffrey calls them) of the possible states that might be brought about by the action.

Within a very general framework such as Jeffrey's, a decision problem is given, not by a probability-utility matrix of the form displayed in Table 1.3, but by one having the form given by Table 1.5, in which $P_\alpha(S)$ is the probability on state S induced by action α and $U(\alpha, S)$ is the utility of the consequence action α in state S. The principle of maximisation of expected benefit then requires choice of the action α that maximises the quantity:

$$V(\alpha) = \sum_{i=1}^{n} P_\alpha(S_i) \cdot U(\alpha, S_i)$$

When states are probabilistically independent of acts and consequences desirabilistically independent of states, this just reduces to Savage's theory, while if the former holds but not the latter, it reduces to the state-dependent expected utility theory proposed by Karni (1985) and others. Jeffrey took the P_α to equal $P(\cdot|\alpha)$, a point on which contemporary causal decision theorist disagree. We will return to this issue later.

Two features of this treatment are noteworthy. Firstly, it is not required that the consequence of an action in each state be known in order that a deci-

[2] A criticism which he delivers against Ramsey as well.

sion be made. All that is required is that the agent have probabilities for the consequences given the choice of act. This relaxes the first constraint on the applicability of Savage's theory. Secondly, it is no longer required that the states of the world be probabilistically independent of the available actions. On the contrary, as Jeffrey sees it, actions matter precisely because they influence the probabilities of states (if you like, the consequences of acting *are* changed probabilities of states). This dispenses with the third constraint on the applicability of Savage's theory.

The fact that Jeffrey's theory imposes much weaker requirements on the framing of decision problems provides strong grounds for preferring his framework to Savage's for developing a decision theory with human face. There are other advantages too, such as the simplicity and flexibility of a working with sets of propositions and the fact that the foundational representation theorems for his theory require much weaker assumptions about rational preference. But there is a downside to this flexibility and indeed controversy over whether his framework (in particular his conception of acts) captures all relevant features of decision problems. So in the next part of the book I will develop a version of Bayesian decision theory that follows Jeffrey's in defining degrees of belief and desire on a common Boolean algebra of prospects (his propositions). But I will show how it is possible to extend the set of prospects in a way that allows for the re-introduction of Savage-style acts and a formulation of a state-dependent version of his theory. This richer theory is the one that I will defend as giving the best account of a certain kind of ideal rational agency.

Chapter 2

Rationality

2.1 Moderate Humeanism

The maximisation of expected utility hypothesis brings together two separate claims. The first concerns what rationality requires of the relation between the agent's preferences between different prospects and her beliefs and desires. Stripped of mathematical baggage, the claim can be expressed as follows:

Rationality Hypothesis Rationality requires of an agent that she prefer one prospect over another if and only if the expectation of benefit conditional on the truth of the former is greater than the expectation of benefit conditional on the truth of the latter, relative to her degrees of belief and desire.

The Rationality Hypothesis is generally taken to express nothing more than a consistency requirement on the agent's preferences, akin to the requirements that logic places on her beliefs. Consistency requirements are purely formal in nature and place no substantial constraints on the content of any preference, belief or desire taken in isolation. Moreover the constraints that it places on sets of such preferences, beliefs and desires are not such as to rule out many that we might be inclined to regard as defective in some way; for instance because they are immoral, self-destructive or just plain ill-considered.

In ordinary talk we tend to be more demanding and speak of beliefs as irrational, even if they are consistent, because they fail to meet some other standard of adequacy. For instance, we might be inclined to criticise someone for not taking into account all available evidence or for failing to give the long term consequences of their choices sufficient weight. Such talk, it seems to me, runs together two types of requirements that are best kept separate. One is the requirement that we recognise all the available evidence and that we give appropriate weight to all the possible consequences of our actions; the other that our beliefs be consistent with all the evidence that we recognise and that our preferences for actions be consistent with the weight that we give to each of their possible consequences. The former is a requirement that our judgements respond in an adequate way to the world as it is presented to us; the latter is a consistency requirement.

Let us call requirements of the first kind, *internal* requirements and of the second, *external* requirements. A very demanding example of an external requirement is that the agent form only true beliefs and correct desires. A weaker one would be that we form only beliefs and desires that we have reason to form. Yet another that beliefs and desires be formed by a reliable process, one that tends to produce beliefs that are true and desires that promote some benefit. Whatever their merits, it is doubtful that someone who fails to satisfy any of these requirements is irrational. Possibly deficient in some way, but not irrational. The fact that the public transport is not running gives me reason to take my car to work. But if I don't know about the state of the public transport, then my choosing to use it is not irrational. And if I do know about it, then my irrationality stems from the violation of the consistency requirement encoded in the expected utility hypothesis: I fail to do what seems best by my own lights. Could it not be said to be irrational not to find out about the state of public transport? Only if I have beliefs (such as that the system is unreliable) that makes it rational by the lights of the maximisation hypothesis to seek more information.

The view that rationality places only formal constraints on our attitudes contrasts with the view that beliefs and desires that are contrary to reason are irrational. A belief or desire is contrary to reason when there is a decisive reason not have it. Parfit (2013) offers the example of the person who cares about his future pains and pleasure except when they will occur on a Tuesday; not because he believes that the pains and pleasures on Tuesday will be less painful or pleasurable, but simply because of the day on which they occur. As he prefers a pain on Tuesday to one on Wednesday even though he has no reason to, his preferences are, according to Parfit, irrational. Now there is no doubting the odd nature of this person's preferences. But are they irrational? That depends on whether he recognises that pains on Tuesday should weigh equally as pains on Wednesday. If he does, then he is being inconsistent. If not then he is no more irrational than the person who prefers chocolate to strawberry icecream, not because of the difference in taste but simply because it's chocolate rather than strawberry. Being Tuesday or being chocolate is, for these people, an intrinsic reason for preference. They may be mistaken, but being wrong is not the same as being irrational.[1]

John Broome (1999) gives the name 'moderate Humeanism' to the view expressed here—that rationality only constrains our attitudes indirectly by disallowing certain combinations of beliefs, desires and preferences—and argues that it is not a viable position to hold. His argument is that consistency conditions cannot constrain our attitudes at all unless rationality sets some limits to what kinds of distinctions between prospects can support different attitudes to them. There is something right and something wrong about this claim. It is true that without some requirements of indifference, as Broome calls them, consistency cannot constrain our attitudes. But these requirements of indifference do not have to be requirements of *rationality*. Substantial value commitments will do the job, e.g. to treating people impartially or to taking care of oneself. It is no

[1] All of this underscores the extent to which the project of determining the rationality requirements on agents' attitudes is a truly modest one. Consistency is an important property of attitudes, but not *that* important. It is no doubt sometimes better to be warmhearted or generous than to be consistent and a thorough-going consequentialist will not attempt to maximise consistency to the exclusion of all else.

part of the theory of rationality that we should have one value commitment or another, but once we do then formal consistency conditions such as transitivity will work to constrain preferences in all kinds of ways.

The point is more general. To say that rationality qua consistency cannot arbitrate between different sets of beliefs and preferences is not to say that we have no grounds upon which to do so. We certainly can criticise someone for failing to take account of the reasons that they have for preferring one prospect to another just as we can criticise someone for failing to attend to the evidence adequately in forming their beliefs. In doing so we might appeal to external requirements on their preferences; to the facts (as we see them) about what is worthy of preference. Such appeals need not involve adopting a value standpoint which the agent rejects. Suppose I am a hedonist and regard the consumption of Cassoulet on a cool evening as the greatest pleasure. Another may criticise me, by saying that I had failed to properly attend to the superior qualities of freshly grilled sea bass served on a warm evening by the sea, perhaps because of a cultural bias or insufficient experimentation. They criticise me, not for possessing inconsistent preferences, but for poor judgement or poor application of my own values. So too might one be criticised for failing to live up to one's moral commitments or to appreciate what they require of one.

Moderate Humeanism should not be confused with other views to which the label 'Humean' has been attached. In particular it does not entail the Humean theory of motivation, according to which belief is never sufficient to motivate action, requiring the presence of desire. Whether this theory is correct or not is largely a matter of empirical psychology, something on which a theory of rationality cannot legislate. Nor does moderate Humeanism entail either moral noncognitivism or moral subjectivism, both sometimes attributed to the 'Humean' view (more on this later). Finally, it is frequently said to be part of the Humean view that not only are preferences and desires distinct from beliefs, but beliefs do not constrain preferences or desires at all (or vice versa). In his recent book, Ken Binmore (2008) calls this Aesop's Principle and gives the following statement of it:

Aesop's Principle: Preferences, beliefs and assessments of what is feasible, should all be independent of each other.

The conviction that Aesop's principle is fundamental to decision theory seems wide-spread. But, as Binmore himself notes, it is easy enough to find objections to the principle. Indeed not only does decision theory *not* generally require independence of preference from belief, it requires that preferences *be* sensitive to it. What I believe about the weather conditions should influence my preferences over clothing, what I believe about the freshness of the food being served at different restaurants should influence my preferences about which of them to frequent, and so on.

What Binmore really means is that a particular class of preferences are governed by Aesop's Principle, namely fundamental or intrinsic preferences. A preference for one thing over another is intrinsic, according to Binmore, if nothing we can learn would change it.[2] They are thus unconditional in the sense

[2]Binmore says nothing that can happen would change it, but this is too strong. Even intrinsic preferences could be changed by a blow to the head or some other non-informational disturbance to mental states.

that they do not depend on some or other condition being satisfied or, more exactly, on the belief that the condition is satisfied. In contrast instrumental preferences are preferences for prospects that do depend on them being a means to some other good. They are thus conditional on one's beliefs about the kinds of things that make them more or less efficacious as a means. I like to eat at a local Italian restaurant because I expect to get a tasty meal there. Its desirability derives from being instrumental to tasty experiences and is, therefore, conditional on the quality of the cooking not declining and fresh ingredients having been delivered that day. Many prospects are both instrumentally and intrinsically desirable. I take the dog for a walk because it gives both of us the exercise that we need and because I like doing it. If the need for exercise were removed, I would still walk, but less frequently, and not when the weather was foul.

Once we restrict Aesop's principle to intrinsic preferences, it becomes more or less empty, since it seems to be part of the definition of an intrinsic preference that it should satisfy the principle. The substantive issue is whether there are any preferences that we generally hold that are fundamental in the required sense. I don't see any reason to believe that there are. Being wealthy, attractive and in good health are no doubt all things that we might desire under a wide range of circumstances, but not in circumstances when these arouse such envy that others will seek to kill us or when they are brought about at great suffering to ourselves or others. Even rather basic preferences such as for chocolate over strawberry ice-cream are contingent on beliefs. In any case, the important point is that the maximisation principle itself does not require that there be prospects that are intrinsically desirable. It requires our preferences for actions to be consistent with the value we attach to its consequences, but not that the value that we attach to these consequences be unconditional, non-revisable or fundamental.

2.2 The Choice Principle

The Rationality hypothesis alone does not say anything about what agents should or will do. For this a second claim, connecting preference to choice, is required.

Choice Principle: Of the options available to an agent, she *should/will* select the one that she most prefers.

The Choice Principle either states a descriptive or a normative claim, which together with the Rationality hypothesis yields descriptive and normative versions of expected utility theory. To asses either, it is essential to be clear about the interpretation of the maximisation hypothesis being worked with. In fact, at least two quite different uses of expected utility theory have been doing the rounds for some time. In one usage, maximisation of expected utility means something like doing what is in one's best interest, be this a matter of experiencing pleasure and avoiding pain, or of acquiring wealth, power and reputation, or of having a high level of welfare or wellbeing. In another usage, maximising expected utility is a matter of doing what one thinks is best, all things considered, in the light of one's beliefs and preferences. The two usages are

quite distinct. The act that one thinks best may not be one that is in one's self-interest (e.g. lending money to an unreliable friend) and vice versa.

These two usages are instances of the two broad classes of interpretations of the utilities and probabilities figuring in the statement of the maximisation hypothesis that are to be found in the literature. On *empirical* interpretations, probabilities and utilities are features of the world relevant to the agent's decision. In debates on probability, for instance, the view that they are long-run frequencies, that they are propensities of physical systems and that they are objective chances of events all belong to this group. Important objective interpretations of utility include the view that they are hedonic states, that they are degrees of preference satisfaction and that they are measures of wellbeing or welfare.

On *judgemental* interpretations, on the other hand, probabilities and utilities are features of judgements or states of mind. Subjective Bayesians, for instance, view them as measures, respectively of the decision maker's degrees of belief in the various possible states of the world and degrees of preference or desire for the possible consequences. Bayesianism is the predominant view in contemporary decision theory, but other judgemental views have been important in probability theory: in particular, the 'logical' interpretation of conditional probability as a measure of degree of confirmation or entailment between propositions.

For present purposes the most important distinction is between interpretations, such as the empirical and logical ones, which imply that probability and/or utility is something objective ('in the world') and hence something that one can be right or wrong about, and interpretations, such as the Bayesian one, which view them as features of subjective judgement ('in the head'). In principle, a subjective interpretation of utility could be combined with an objective interpretation of probability and vice versa. So even this crude subjective-objective distinction allows for four different values to be attached to an action, at least three of which have figured prominently in applications of decision theory. Mainstream Bayesian decision theory is doubly subjective, but in the von Neumann and Morgenstern theory probabilities are objective and in the social ethics of Harsanyi and others utilities (qua welfare) are objective. There has been much debate over the correct interpretation of both probability and of utility, but I see no reason to think that there should only be one correct construal of either notion. It is better to regard probability and utility as formal notions which can in principle admit of more than one interpretation and debate the appropriateness or usefulness of each for particular applications.

In this regard, three questions are of immediate importance. Which interpretations are appropriate to the Rationality hypothesis? Which interpretations explicate the role played by the Principle of Choice in the description and explanation of action? And which support its application to normative problems of choice? My earlier claim that rationality is a matter of consistency in one's judgements, not of right relation to features of the world, commits me to a subjective interpretation of the first—the Rationality hypothesis. But before exploring the exact nature of such a subjective interpretation, let me make a few comments about the other two questions, without claiming thereby to do them proper justice.

Both subjective and objective interpretations of the maximisation hypothesis are often employed in applications of expected utility theory to the explanation of human behaviour. This has been the cause of a good deal of confusion and

misdirected critical discussion. The claim that agents maximise the objective expectation of utility is clearly much stronger than the claim that they maximise their subjective expectation of it. Moreover, there is little doubt that the former claim is false, since false belief is an important causal factor in people's choices. This doesn't mean that these applications are of no explanatory use. There may be contexts in which the hypothesis that agents maximise some kind of objective expected utility (e.g. self-interest) yields good approximations of actual behaviour, perhaps because relevant information is easily accessible or because agents have opportunities to correct their judgements. But in these cases the 'deeper' explanation resides in the subjective version of the maximisation hypothesis which has the resources both to explain why agents sometimes make the choices that cohere with objective criteria and why sometimes they do not. The problem for the subjective version, on the other hand, is that its claims are notoriously difficult to test. Its critics, consequently, are divided into those who claim that it is unfalsifiable and therefore unscientific and those who claim that it has been falsified.

The normative hypothesis also makes quite different claims, depending on the interpretation given to expected utility. In this case however it is the subjective version that faces the most difficulty. The objection is obvious. Why should subjective expected utility serve as the measure of the choice-worthiness of an action and not, for instance, welfare or moral worth? To put the question more bluntly: on many standards of what it is best for someone to do, the best action will not be the one that maximises subjective expected utility. When someone has false beliefs, choice in accordance with expected utility may lead to a very poor outcome for them, e.g. when they mistake the vinegar for wine and drink it. If this is the case, then surely they should not pick the option that maximises subjective expected utility over an alternative that will in fact deliver a better outcome, even by their own value standards.

The claim that agents should maximise the objective expectation of utility is less vulnerable to this objection. Still it might be argued that we should pick the option that will in fact have the best outcome, not the option with the most favourable expectations. The person who picks the lottery ticket with the greatest expected pay-off may well find themselves wishing that they had picked differently. 'I should have chosen the other ticket' is a reasonable thing to say when it turns out that the alternative was the winning ticket. We might say, parroting Frank Ramsey (1990), that if asked what option we should choose, we should answer "the one that will have the best outcome". But this violates the dictum of 'ought implies can': in situations of objective uncertainty, we simply cannot know what the outcome will in fact be. The objective probabilities characterising a lottery express the limits of humanly attainable knowledge. The most that we can be asked to do to is to make the best attainable judgements and decide consistently on the basis of those.

But why stop there? If the 'ought implies can' dictum can be used to defend objective maximisation, it can also be used in defence of subjective maximisation. For at the time of making a decision, knowledge of the true probabilities and utilities may be impossible. We are where we are, with the judgements that we have arrived at, and at the moment when the decision must be made the best that we can do is act consistently on the basis of those judgements. From the agent's own perspective to maximise subjective expected utility just *is* to do what is best (on their estimation). This is not inconsistent with the

possibility that others will have different views about what that agent should do nor with the claim that they should have done more to improve their opinions. When I say to someone 'you should do x' I am saying something like 'If I were the decision-maker I should do x'. When I am better informed than them, they would do well to listen to what I recommend; indeed consistency will demand it of them if they actually believe that I am better informed. No-one is more expert than the Truth, so when it speaks, we should all listen. But this is not the same as saying we should hear it.

2.3 Subjectivism

The view that the probabilities and utilities figuring in the maximisation hypothesis are the agent's degrees of belief and desire is the predominant one in contemporary decision theory. But this view is only partially correct. To get a handle on what is at stake let us look at how the quantities occurring in a probability-utility matrix should be interpreted. Here a slightly different interpretation is required depending on whether we view the issue from the perspective of the decision maker or from that of an observer. If a decision maker wants to evaluate an action in the manner suggested by the maximisation hypothesis, she must arrive at judgements about the relative likelihoods of the various possible states of the world and desirability of the various possible consequences of her action. The probabilities and utilities figuring in the calculation of the expected utilities of actions are thus her *judgements*. An assignment of probability x to state S, for example is a judgement that S is likely to degree x to be the actual state of affairs. Similarly an assignment of y to consequence C is a judgement that C is desirable to degree y.

When an observer models the choice confronting the decision maker she can either do so from her own point of view or from that of the decision maker. In the former case, she is adopting a first person perspective on the choice problem and so once again the probabilities and utilities she employs are her judgements of likelihood and desirability of truth. In the latter, the probabilities and utilities she writes down are (her estimates of) the decision maker's degrees of belief and preference or desire. What makes it appropriate to model an agent in this way—by imputation of degrees of belief and desire of a particular magnitude to the agent—is the fact that the attributed states play the right kind of causal role in the production of her actions. By right kind of role I mean that they explain, on the assumption that she maximises subjective expected utility, the pattern of choices that she makes. To play this role it is not essential that they be formed as a result of a conscious judgement on the part of the agent. They could, for instance, be part of the agent's cultural or biological inheritance encoded as behavioural dispositions. So it is possible to model the decision making of creatures in terms of maximisation of expected utility even if these creatures don't themselves have the cognitive resources to model the choice problem for themselves. That is, we can adopt a third-person perspective on the utility maximising actions of agents who do not themselves have the corresponding first-person perspective on the decision problem. When we explain an animal's food choices, for instance, we can offer an explanation in terms of its beliefs about what plants are fit to eat, even if the animal doesn't have a concept of 'fit to eat'. But such an explanation is often less satisfactory than one which

is couched in terms of the concepts recognised by the agents themselves. If the animal prefers green foods over red ones, and green foods happen to be those that are fit to eat, then it will be possible to explain her choices in terms of what is fit to eat (that is, such an explanation will cohere with the pattern of her choices) even when it is her colour-of-food judgements that are causally responsible for the development of her choice dispositions.[3]

Both of these two essentially subjective interpretations of probability and utility, as judgements and as mental states, offer an appropriate interpretation of the maximisation hypothesis as a claim about rationality. On the judgement interpretation it says that rationality requires of agents that they judge actions to be desirable to the degree that they can be expected to have desirable consequences, given how likely they judge the possible states of the world to be and how desirable they judge the possible consequences. Similarly on the mental state interpretation, the hypothesis says that an agent is rational only if the value she attaches to each action is its expected desirability, relative to her degrees of belief and desire.

The two interpretations are quite closely related, and it is perhaps not surprising that they are not clearly separated in Bayesian decision theory, the predominant subjectivist view. Indeed, since making a judgement leads to forming a corresponding belief or desire it is rather natural to think of the judgement interpretation as just a special case of the belief-desire one. But it is a mistake to do so: although the latter view is the correct one to take in regard to modelling other agents' decisions, it is not satisfactory for first-person normative applications. When we try to make up our mind about what action to perform by attaching utilities to consequences and probabilities to states we are not aiming to describe our own attitudes but to determine what the relevant features of the decision problem are: whether some condition is likely to hold, whether one consequence is preferable to another, and so on. We are making a judgement about the *world*, not about ourselves, and it is accuracy with regard to the former not the latter that concerns us. For this reason, the right interpretation of notions like probability and utility, in these applications, is as judgements of a particular kind.

I stress this rather subtle distinction because of its implications for a related issue. Many Bayesians not only adopt a subjective interpretation of probabilities and utilities but also deny the existence of objective probabilities and utilities of any kind—a view that is known as Subjectivism. Subjectivism has had a number of famous advocates, including De Finetti, Savage and Jeffrey, but although the arguments for and against their position are quite well known (in probability theory at least) there has been little recognition of an important ambiguity in it. When subjectivists hold that probabilities and utilities are 'in the head' rather than the world, they can mean two quite distinct things. On one (cognitivist) interpretation, a statement such as 'the probability of rain is one half' is true or false depending on what the agent believes. That is to say, probability and desirability statements are truth-susceptible propositions about the mental states of agents. Hence both refer to that part of the world occupied by the agent's head. On a second (expressivist) interpretation, such statements do not make descriptive claims at all. Rather they express an evaluative judgement by

[3] I take this to be the heart of the claim of hermeneutic philosophies that explanation of human action requires understanding, glossed here as identifying the categories that the agent herself uses to formulate the decision problem that she faces.

the agent that is not susceptible to truth or falsity. Judgemental probabilities are not in the (material) world at all.

Both views are tenable. When a probability statement is made in the context of describing, from an observer's point of view, the attitudinal state of an agent, then the statement should be read as a description of the agent that is either true or false of her. On the other hand, when the agent herself makes such a statement, say in the context of thinking through a decision problem, then she should not be read as describing her own state of mind, but rather as making her mind up by reaching an opinion on features of her environment.

Many decision theorists not only fail to recognise these distinct possibilities but, perhaps as a consequence, adopt a rather extreme subjectivist position on value; a view that I will dub Ethical Subjectivism. Just as expressivist subjectivists about probability argue that an assertion about probability is an expression of partial belief not a claim about a feature of the world, Ethical Subjectivists about value view desirability statements as expressions of preference rather than assertions about some objective value. But they go a step further. Subjectivists about probability typically do not deny that there are objective features of the world that are tracked by probability judgements, just that these features are themselves probabilities. They take probability judgements to be subjective judgements on objective, but non-probabilistic facts. Ethical Subjectivists, on the other hand, not only deny that there are objective utilities, but also that any objective feature of the world at all is tracked by utilities. Utility judgements, on this view, are not subjective judgements of the degree to which the world conforms to one or more objective value standard, but (bare) expressions of the agent's subjective tastes or emotions.

This is a much stronger view than the kind of moderate Humeanism that I was defending earlier on. Moderate Humeans hold that the only constraints that rationality places on desirability judgements are formal ones. This, I argued, was consistent with the view that these judgements may be better or worse with respect to satisfaction of external requirements of one kind or another. Ethical Subjectivism implies a denial of this latter view, since the only requirements it recognises are those of consistency. There is no reason why a subjectivist, even of the expressivist variety, should accept this view. One may consider that utilities express a judgement on the part of the agent and at the same time deem that this judgement can be more or less adequate in the extent to which it coheres with, or tracks, some kind objective value. All that an subjectivist about utility needs to deny is that a utility judgement is a belief that something has a certain objective utility. But utility judgements can be subjective judgements concerning objective properties of the world, so long as these properties are not themselves utilities, just as probability judgements can be subjective judgements on the facts without these facts having the structure of probabilities.

Finally let me emphasise that the adoption of one or another subjective interpretation of the Rationality Hypothesis does not entail a commitment to subjectivism in any of its forms. There are conceptions of objective probability (e.g. as frequencies or chances) and of objective desirability or utility (e.g. as wellbeing or goodness) that play an important role in decision theory; both as properties of states of affairs that agent's do in fact care about and perhaps as properties they should take into consideration. In particular, it is hard to deny that we do experience some uncertainty as objective. This fact can, and should, be accommodated by decision theory, even one which adopts a subjective

interpretation of the main decision variables.

Chapter 3

Uncertainty

Uncertainty is a pervasive feature of human life and almost all our decisions must be made without certainty about what the consequences of our actions will be for ourselves or others. Human attitudes such as hope, fear and even regret depend on it. Despite this, philosophy has given little attention to uncertainty, largely treating it as just lack of certainty (apparently the real interest). In both the mathematical and empirical sciences, on the other hand, the emphasis has been on the development of techniques to manage it. So central has the concept of probability become to this enterprise that it has come to seem as if uncertainty was nothing other than the flipside of probability.

The concept of probability emerged surprisingly late in human history, in the 17th century work of Pascal and Fermat (see Hacking (2006)). In this work, and much that followed, probability was conceived both in terms of the stochastic properties of chance processes, such as dice roles and card deals, and the properties of beliefs about events regarding which full knowledge was lacking. In time this hardened into a distinction between two different forms of uncertainty: objective or *aleatory* uncertainty, which derives from features of the world (indeterminacy, randomness) and subjective or *epistemic* uncertainty, which derives from lack of information about it. In modern decision theory, there is a dominant theory of how each form of uncertainty should be quantified (in both cases, by a probability measure) and of how, so quantified, it should weigh in the evaluation of actions. For situations of objective uncertainty (or risk, as it typically called) decision theorists look to the version of expected utility theory originally due to John von Neumann and Oscar Morgenstern von Neumann & Morgenstern (2007/1944), while for those characterised by epistemic uncertainty they look to subjective expected utility theory, whose classic statement is to be found in the work of Leonard Savage Savage (1974/1954).

This distinction between risk and epistemic uncertainty, important and useful though it may be, does not remotely do justice to the variety of forms and degrees of uncertainty relevant to decision making. Firstly, epistemic uncertainty comes in different degrees of severity that derive from differences in the quantity and quality of information that we hold. There is a significant difference, for instance, between being unsure about when someone will arrive because one lacks precise information about their time of departure, traffic conditions, and so on, and having absolutely no idea when they will arrive because you don't know when or whether they have left, whether they are walking or

driving or indeed whether they even intend to come. In the former case, the information one holds is such as to make it possible to assign reasonable probabilities to the person arriving within various time intervals. In the latter, one has no basis at all for assigning probabilities, a situation of radical uncertainty or *ignorance*. It may be rare for us to be totally ignorant, but situations of partial ignorance, or *ambiguity*, in which the decision maker is unable to assign determinate probabilities to all relevant contingencies, are both common and important.

Secondly, according to some critics, Bayesian theory fails to distinguish between the different levels of confidence we might have, or have reason to have, in our probability judgements. Compare a situation in which we are presented with a coin about which we know nothing to one in which we are allowed to conduct lengthy trials with it. In both situations we might ascribe probability one-half to it landing heads on the next toss: in the first case for reasons of symmetry, in the second because the frequency of heads in the trials was roughly 50%. It seems reasonable however to say that our probability ascriptions are more reliable in the second case than the first and hence that we should feel more confident in them.

Both of these issues will be discussed in detail in the second half of the book. The focus on my concern now will be a third issue. Decision makers confront uncertainty not just concerning what is the case (empirical or factual uncertainty), but also what should be the case (evaluative uncertainty), what could be the case (modal uncertainty) and what would be the case if we were to make an intervention of some kind (option uncertainty). Almost all discussion of uncertainty is directed at the first of these. The others are just as important however and so I shall attempt in this chapter to say something about them, exploring the question of how they should be captured and whether they can be reduced to a form of empirical uncertainty.

3.1 Evaluative Uncertainty

Although the distinction between certainty and uncertainty is typically used only to characterise the agent's state of knowledge of the world, it is equally important to distinguish cases in which consequences have known, or given, objective values and those in which these values are either unknown and the decision maker must rely on subjective evaluations of them, or do not exist and the decision maker must construct them. The possibility of such evaluative uncertainty is typically ignored by decision theorists, because of their (often unconscious) attachment to Ethical Subjectivism, the aforementioned view that values are determined by the agent's subjective preferences. If this view were correct, talk of evaluative uncertainty would be misleading as one is not normally uncertain about what one's own judgement on something is (just what it should be). Indeed it makes questions such as 'What utility should I attach to this outcome' seem barely intelligible. If a prospect's value for an agent is determined by her preferences, she cannot be right or wrong about what value to attach to them; nor can her preferences be criticised on grounds of their failure to adequately reflect one value or another.

There are however at least two ways in which one can be uncertain about the value to attach to a particular consequence or whether one consequence is

preferable to another. Firstly one may be uncertain about the factual properties of the consequence in question. If the latest Porsche model is the prize in a lottery, one may be unsure as to how fast it goes, how safe it is, how comfortable and so on. This is uncertainty of the ordinary factual kind and, if one wishes, it can be 'transferred' from the consequence to the state of the world by making the description of the consequence more detailed. For example, the outcome of the lottery may be regarded as having one of several possible consequences, each an instantiation of the schema 'Win a car with such and such speed, such and such safety features and of such and such comfort', with the actual consequence of winning depending on the uncertain state of the world.

Secondly one can be unsure as to the value of a consequence, not because of uncertainty about its factual properties, but because of uncertainty about how valuable these properties are. One may know all the specifications, technical or otherwise, of the latest Porsche and Ferrari models, so that they can be compared on every dimension, but be unsure whether speed matters more than safety or comfort. Once all factual uncertainty has been stripped from a consequence by detailed description of its features, one is left with pure value uncertainty of this kind.

When we assume that values are given, we take this uncertainty to have been resolved in some way. This could be because we assume that there is a fact of the matter as to how good a consequence is or as to whether one outcome is better than another, a fact that would be detailed by the true axiology. But it could also be because the description of the decision problem itself comes with values 'built-in'. For instance, in a problem involving a decision between two courses of medical treatment, it may be that a limited number of value considerations apply in the assessment of these treatments: number of patients saved, amount of discomfort caused, and so on. The decision theorist will be expected in such circumstances to apply only the relevant values to the assessment of the options, and to set aside any other considerations that he or she might 'subjectively' consider to be of importance.

In many situations, however, values are not given in any of these ways and the agent may be uncertain as to the value she should attach to the relevant prospects. She may, for this reason, also be willing to revise her evaluations in the face of new considerations or the criticism of others. These facts would seem to render Ethical Subjectivism unsustainable. But the Ethical Subjectivist can insist that an agent cannot be wrong about what value to attach to fully specified consequences, and point out that it suffices that there be factual uncertainty for us to be unsure about the desirability of any less than fully specified prospect. Since in practice complete specification is impossible, this means that evaluative uncertainty of the kind that derives from factual uncertainty will be ubiquitous. Furthermore, since we may hold false beliefs our ethical judgements on incompletely specified prospects are certainly criticisable. All that is ruled out is pure value uncertainty.

Other views take the possibility of value uncertainty more seriously. There are three that I will mention here. The first is that evaluative uncertainty is just ordinary uncertainty about agents' tastes or, more generally, about the features of agents that are relevant to the utility of the option. The thought is this. What value an agent will assign to a commodity depends not just on features of the commodity itself (the speed, safety and comfort of the cars) but also on features of the consumer: their likes and dislikes, their capacities

(for instance, their driving skills) and their needs. One can be just as uncertain about the latter as the former. This view has some application to decisions with consequences for different people or ones far in the future and where the value we attach to these consequences depends of the attitudes that the different people, or our future selves, takes to them. But it is not plausible as a general account of value uncertainty. When we are uncertain about whether it is more important to help a friend or to further one's own interests, the difficulty that we have in deciding the question stems not from the fact that we don't know what we in fact prefer but that we don't know what we *should* prefer. Indeed I doubt that in such cases there really is anything like a set of pre-given preferences waiting to be discovered. An intermediate case would be trying to decide whether to take up the violin or fencing. Can the problem be described as trying to work out what one's tastes are? I think not. One's tastes are likely to be shaped by the decision itself, for in pursuing the violin one will learn to appreciate one set of skills, in taking up fencing one will learn to appreciate another.

On a second cognitivist view, what I am calling value uncertainty is just ordinary uncertainty about normative facts. The uncertainty I experience about whether to help my friend is uncertainty about whether it is in fact good to help one's friend or whether it is true that it is better to help one's friend than further one's own interests. So, on this view the difference between uncertainty about whether it will rain and about whether it is good that it rains is to be located in the type of proposition about which one is uncertain, not in the nature of this uncertainty. It is an open question whether this position is consistent with Bayesian decision theory. David Lewis (1988)Lewis (1996) famously argued that it is not, but others (Broome (1991), Bradley & Stefansson (2016)) that Lewis' argument is mistaken.

Both these cognitivist views treat value uncertainty as a kind of factual uncertainty, differing only with regard to the kinds of facts that they countenance and consider relevant. They are in that sense reductive views. The last of the views I want to consider holds that evaluative uncertainty is different in kind from factual uncertainty and is directly expressed in utility judgements, rather than in second-order judgements about tastes or first-order probabilities judgements about normative facts. Making this precise requires some care. Utility judgements are like probability judgements in that they are judgements about the world (and not just expressions of the agent's mental state). But while we can say that one's probability for rain tomorrow, say, reflects the degree to which one is uncertain as to whether it will rain, it is not the case that one's utility for rain expresses the degree to which one is uncertain as to whether it is true that it is good that it rains. Rather it expresses one's uncertainty as to how good it would be if it rained. On the reductive views, once we know all the facts—about what will happen when it rains, how much people like getting wet, and so on—all uncertainty is removed and the value of rain is fully determined by either the relevant normative facts or by the agent's subjective degrees of desire for rain, given the facts. On the non-reductive view, even when we know all the facts we can be unsure as to how desirable rain is, given the facts.

It will not matter to this book which of these views of evaluative uncertainty is adopted, so long as it is consistent with Bayesian decision theory. On any such view, evaluative uncertainty is captured or measured by the agent's value function. Since evaluations generally depend on the facts, it follows that a value function that adequately represents an agent's state of evaluative uncertainty

	States of the world		
Options	S_1	...	S_n
α	$\{a_1^1, ..., a_1^j\}$	...	$\{a_n^1, ..., a_n^j\}$
...	...	...	...
γ	$\{c_1^1, ..., c_1^j\}$	...	$\{c_n^1, ..., c_n^j\}$

Table 3.1: State-Consequence Correspondence

must be revisable in the face of new factual information (and potentially new value experiences as well). This desideratum figures prominently in the choice of value function that I make in the second part of the book and which differentiates it from the utility functions standardly employed in decision theory.

3.2 Option uncertainty

In the state-consequence representation of a decision problem that we started with, actions were associated with definite consequences, one for each state of the world. But in real decision problems we are often unsure about the relationship between actions, worlds and consequences, either because we do not know what consequence follows in each possible state of the world from a choice of action, or because we don't know what state of the world is sufficient, for a given action, to bring about that consequence. For instance, we may be uncertain as to whether taking an umbrella will certainly have the consequence of keeping us dry in the event of rain. Perhaps the umbrella has holes, or the wind will blow it inside out or the rain will be blown in from the sides.

We can put this difficulty in slightly different terms. A possible action may be defined by a particular mapping from states to consequences. Then no uncertainty about the mapping itself can arise. But what we will then be unsure about is which actions are actually available to us i.e. which of the various hypothetical actions are real options. Whether we describe the problem in these terms—as uncertainty about what options we have—or as uncertainty about the consequences, in each state of the world, of exercising any of the options we know we have, is of little substance, and I shall use the same term—option uncertainty—to denote both.

Option uncertainty arises when we are unsure about what would happen if we were to act in some way, or perform some kind of intervention in, or manipulation of, our environment. When an agent faces option uncertainty, she cannot represent her decision problem with the simple state-consequence matrix represented by Table 1. But she can do something quite similar by replacing the fine-grained consequences that play the role of Savage's 'sure experiences of the deciding person', with *sets* of such fine-grained consequences—intuitively the set of consequences the agent regards as possible given the act and state in question. This is exhibited schematically in Table 3.1 in which each act γ is represented as a function from each state S_i to a set of associated possible consequences $\{c_1^1, ..., c_1^j\}$. The larger the sets of possible consequences the greater the option uncertainty facing the agent.

There are three strategies that can be pursued in handling option uncer-

	States of the world		
Options	$S_1(\alpha^i)$	...	$S_m(\alpha^i)$
α^1	C_1^1	...	C_m^1
...	...	...	...
α^n	C_1^n	...	C_m^n

Table 3.2: State Functions

tainty. The first is to try to reduce or transform option uncertainty into empirical uncertainty about the state of the world. The second is to reduce it to evaluative uncertainty. And the third is to treat it as a *sui generis* form of uncertainty. Let's consider each turn.

Reduction to Empirical Uncertainty: Decision theorists typically attempt to reduce option uncertainty to uncertainty about the state of the world by refining their description of the states until all contingencies are taken care of. They will regard a state of the world as insufficiently described by the absence or presence of rain, for instance, and argue that one needs to specify the speed and direction of the wind, the quality of the umbrella, and so forth. There at least two reasons why this reductive strategy will not work on all occasions. Firstly because, according to our best scientific theories, the world is not purely deterministic. When the conditions under which a coin is tossed do not determine whether a coin will land heads or tails, for instance, the act of tossing the coin does not have a predictable consequence in each state of the world. Secondly, even if we are in a purely deterministic set-up, it may be subjectively impossible for the decision maker to conceive of and then weigh up all the relevant contingencies or to provide descriptions of the states of the world that are sufficiently fine-grained as to ensure that a particular consequence is certain to follow, in each state, from the choice of any of the options open to them. And even if one could envisage all the possibilities, one may simply not know what state of the world is sufficient for the act of taking an umbrella to keep me dry.

To get around these difficulties the reductionist can make a two-pronged attack. To handle objective indeterminacy, she can allow consequences to be objective probability distributions (lotteries) over outcomes, and apply von Neumann and Morgenstern's theory to give a measure of their utility. And to handle enumerability-of-states problems, she can draw on descriptions of the states of the world that identify the set of conditions sufficient for the determination of the consequence, given the performance of the action, without actually listing the conditions. For instance, she can turn Savage's theory around and take actions and consequences as the primitives and then define states of the world as consequence-valued functions ranging over actions. This would lead to a decision matrix of the kind exhibited in Table 3.2, in which each $S_j(\alpha^i)$ denotes the state that maps actions α^i to consequences C_j^i.

This descriptive strategy has some notable advocates. Lewis (1981), for instance, treats states as 'dependency hypotheses', these being maximally specific propositions about how consequences depend causally on acts. Similarly, Stalnaker (1981) suggests that a state of the world be denoted by a conjunction of

conditional sentences of the form 'If action A were performed then consequence C would follow; if action A' were performed then consequence C' would follow; if ... '. In this way option uncertainty is transformed into a particular kind of state uncertainty, namely uncertainty as to the true mapping from actions to consequences or as to the truth of a dependency hypothesis or particular conjunction of conditionals.

Reduction to Evaluative Uncertainty: A second strategy for dealing with option uncertainty is to coarsen the description of the consequences to the degree necessary to ensure that we can be certain it will follow from the exercise of an option in a particular state. (Savage, 1974/1954, p. 84), for instance, acknowledges the need for "acts with actually uncertain consequences to play the role of sure consequences in typical isolated decision situations". Formally this amounts to treating the sets of possible consequences associated with an action occurring in Table 3.1 as single coarse-grained consequences and giving them a utility value. Pursuit of this strategy converts option uncertainty, not into ordinary uncertainty about the state of the world, but into uncertainty about the desirability of the consequence as described. We may be sure that the act of taking an umbrella will have the consequence in a rainy state of being able to shield ourselves against the rain by opening the umbrella. But whether this is a good thing or not depends on contingencies that by assumption we are unable to enumerate or identify. How bad it is to get soaked, for instance, depends on how cold the rainwater is and rain temperature may be a variable about whose determinants we know very little. Whatever utility value we assign to the coarse-grained consequence of having an umbrella as rain-protection will embody this uncertainty.

Non-Reduction: The last strategy to consider is to accept the presence of option uncertainty and try and develop a measure of it. We could, for instance, take the path recommended at the end of the last chapter and assign probabilities directly to consequences that depends on the action performed. So instead of trying to enumerate the features of the state of the world that will ensure that I stay dry if I take an umbrella, I simply assess the probability that I will stay dry if I take the umbrella and the probability that I will get wet anyhow (even if I take it). In making these probability judgements, I may well try and conceive of the various contingencies under which staying dry will be a consequence of my action, but I need not be able to conceive of all of them in order to do so. Having done so I may directly represent the decision problem I face in terms of the probabilities and utilities of the various possible consequences induced by each option in the manner of Table 1.5 exhibited in the previous chapter. So while the other strategies led to alternatives to our initial state-consequence matrix representation of a decision problem, this last one offers an alternative to the initial quantitative representation of it.

In recent years a debate between so-called evidential and causal decision theorists has raged as to the nature of these act-dependent probabilities. Evidentialists such as Richard Jeffrey regard the conditional probabilities of possible consequences, given that an action is performed, as giving the correct measure of the uncertainty associated with acting, while causal decision theorists such as James Joyce (1999) argue that what is required is a measure of their prob-

ability under the counterfactual supposition that the action is performed. If the evidentialists are correct then a single probability function suffices not only to measure state uncertainty but option uncertainty as well. But the difficulty that evidential decision theory faces in dealing with Newcomb's paradox and other more homely cases in which probabilistic correlation fails to provide a good guide to causal efficacy suggests that it is not (more on this in the third part of the book). So the issue of how to measure option uncertainty remains open.

3.3 Modal Uncertainty

Empirical uncertainty arises when we are unsure as to what *is* the case and evaluative uncertainty when we are unsure as to what *should* be the case. Modal uncertainty arises when we are unsure as to what is possible or about what *could* be the case: what contingencies might arise, what consequences might follow from our actions and what actions are feasible. In Savage's framework, no modal uncertainty can arise as both the state space and the set of possible consequences are exogenously given. In real decision problems, however, agents must grapple with the possibility of *unforeseen contingencies*: eventualities that even the knowledgeable will fail to take account of. If a decision maker is aware of the possibility that they may not be aware of all relevant contingencies—a state that Walker *et al.* (2011) call 'conscious unawareness'—then they face modal uncertainty.

There are two variants of the problem of unforeseen contingencies that should be distinguished. The first arises when the agent is aware that the states that she can conceive of may be too coarse-grained to capture all decision-relevant considerations. For instance, someone planning where to go on holiday may take into account all factors that seem relevant to their enjoyment of it (costs, climate, cultural amenities and so on) but nonetheless worry that have omitted something which mitigates these factors. As a result she either cannot be sure what the exact consequences of her actions are (i.e. she faces option uncertainty) or, if she can, whether the consequences are sufficiently fine-grained as to capture everything relevant to their value (i.e. she faces evaluative uncertainty).

A second variant of modal uncertainty arises when the agent is aware of the possibility that she has entirely omitted consideration of a possible state or possible consequence. For instance, a business man considering an investment may be unsure as to what new technologies will be available in the future. So he will be unable to exhaustively enumerate all the states determining the return on his investment. In response a catch-all state can be introduced—the 'any other contingencies not yet enumerated' state—thereby eliminating modal uncertainty (superficially at least). But this catch-all state will have a completely unknown consequence, so severe option uncertainty now arises. Furthermore, as we have no way of assigning a probability to this state, severe factual uncertainty is generated.

Modal uncertainty presents Bayesian decision theory with its most difficult challenge. For if we don't know what all the relevant possibilities are, is it really rational to try and optimise relative to those we are aware of? A course of action that is best relative to a limited set of considerations may turn out to be disastrous once the unforeseen ones reveal themselves. We will return to

these issues in the last part of the book.

Chapter 4

Justifying Bayesianism

What reasons do we have for accepting or rejecting Bayesian decision theory? Empirical theories stand or fall on the basis of their ability to handle the facts: above all by the quality of their explanations and the accuracy of their predictions. As an empirical theory of judgement and decision making, subjective expected utility theory has endured a good deal of criticism in the last thirty years or so, with a range of experimental results suggesting that it is a poor predictor of people's behaviour. On the other hand, none of the main rival empirical theories seem to do much better when confronted with data other than that used to generate them; according to some studies they do worse.[1] So while there is every reason to be cautious about the theory's predictive abilities in a wide range of cases, there is as yet no good reason to abandon it entirely.

It is as a normative theory, however, that we are interested in the problem of justifying acceptance of Bayesian decision theory. Although normative theories cannot be refuted by direct appeal to the facts, the general principles of a normative theory can be assessed in a similar way to an empirical one. A scientific theory will have its laws assessed by deriving predictions about concrete cases and then testing to see whether the predictions turn out to be true or not. A normative theory doesn't make predictions, but it will have implications for concrete situations which can be compared with our judgements about what is correct in those cases. A theory of valid inference, for instance, can be tested against concrete instances of inferences that we are inclined (or otherwise) to make; a theory of grammar against sentences that competent speakers find acceptable; and so on.

The fundamental lesson of the Quine-Duhem problem—that the falsity of a scientific hypothesis can rarely, if ever, be deduced from a set of observations—applies equally to normative theories. When general normative principles clash with our judgements regarding a concrete case, all that follows is that we cannot coherently hold onto both. We can revise the principles so that they can accommodate the intuitive judgements or we can revise the judgements themselves. Frequently we can also do neither and instead revise one of the numerous other assumptions that are typically needed in order to draw out the implications of the general principles for concrete cases.

When something must be revised, foundationalists suggest that we retain

[1] See Starmer (2000) for an overview of the empirical evidence.

those principles or judgements which have some kind of special justification and use them as an anvil upon which to beat the rest into shape. Such thinking is plausibly behind the extensive use of axiomatic methods in decision theory, with the role of foundational principles being played by propositions concerning the rationality properties of preferences. Indeed to the question 'What grounds are there for thinking that rationality requires us to maximise subjective expected utility?' decision theorists will typically produce a representation theorem and argue that it shows that the claims of expected utility theory can be derived from 'self-evident' principles of rational preference.

I doubt that the rationality claims about preference that are required for these arguments can bear the full justificatory load typically piled on them. But there is no doubting the importance of representation theorems to the Bayesian enterprise. And as I shall be relying quite heavily on them in the next part of the book it is best to come clean now about their limits as well as their scope. So this chapter will be devoted to looking carefully at these theorems, the assumptions they make and philosophical positions that they are supposed to support.[2] The broad view that I will take is that they are best viewed as moves within a search for a reflective equilibrium, in which general principles and judgements on particular cases are brought in line by systematic refinement of both with the aim of maximising overall coherence, but in which no (class of) proposition plays the role of final arbiter of truth. Since this method can lead to quite different outcomes depending on what choices are made about what to revise, it is perfectly possible that two people pursuing it will end up with different normative theories that achieve equal overall coherence. Nonetheless it seems to me that we cannot get by without relying on this method to a large extent.

4.1 Pragmatism

A representation theorem for a decision theory proves the existence of an isomorphism between two kinds of structures: a class of preferences satisfying a set of conditions and a class of numerical functions with certain properties. A 'ideal-typical' Bayesian representation theorem, for instance, establishes that if an agent's preferences satisfy a particular set of axioms, then these preferences can be numerically represented by a pair of probability and utility functions measuring her degrees of belief and desire, in the sense that one alternative is preferred to another iff the expectation of utility given the former exceeds that of the latter.

The central primitive of these theorems is the notion of preference, reflecting widespread adherence amongst decision theorists to Pragmatism, a view which accords conceptual and methodological priority to preferences over numerical degrees of belief and desires. Methodological priority because preferences, as revealed in the behaviours that they engender, are the empirical basis for attributions of degrees of belief and desire to agents. Conceptual priority, because it is the properties of rational preference that are said to explain the laws of rationality for partial belief and desire.

[2] Despite their centrality, almost all philosophical discussion of representation theorems has been directed at the status of the axioms they invoke, rather than the arguments that they are supposed to support (recent exceptions are Meacham & Weisberg (2011) and Zynda (2000)).

These priority claims should not be confused with those emanating from another commonly held view amongst decision theorists: Behaviourism. Behaviourism accords methodological and conceptual priority to observable behaviour, and particularly choice behaviour, over preference. Methodological priority because observations of choice behaviour are said to furnish the empirical basis for ascriptions of mental attitudes to agents. Conceptual priority because it is the properties of observable behaviour that are said to explain those of preference. These claims are supported by another set of representation theorems, linking choice behaviour to preferences, and which serve to characterise (in the ideal case) the conditions on observed choices necessary and sufficient for their representation by a preference relation having certain properties.

The claim of methodological priority for observable behaviour is rooted in the desire to see the human and social sciences rooted in evidence that is intersubjectively verifiable. It is an entirely defensible position and quite in line with practices in many of these disciplines. But the claim of conceptual priority is much less plausible, and its weakness is evidenced by the failure of attempts by behavourists in many different fields to eliminate mentalistic vocabulary from scientific discourse. Indeed, contrary to the view it expresses, it is the nature of preference and its mental determinants that accounts for the properties of behaviour and not the other way around. It is not surprising therefore that Behaviourism's conceptual priority thesis has fallen into philosophical disrepute.

Pragmatism implies neither of the priority claims of Behaviourism. It is true that many decision theorists share with it a distrust of introspection as a means of determining an agent's mental states. Ramsey, for instance, firmly dismissed the idea that we could introspect our degrees of belief on the grounds that "the beliefs which are held most strongly are often accompanied by practically no feeling at all; no one feels strongly about the things he takes for granted" (Ramsey, 1990/1926, p. 65). Both he and Savage argued instead that our judgements of belief were really about how we would act in different hypothetical circumstances. But neither were thereby dismissing introspection altogether: since we can't tell how we would act in a hypothetical circumstance by direct observation, this information would have to come by introspecting of what we would do. It is introspection of quantitative degrees of belief and desire that they considered unreliable, not introspection in general.

These observations point to a second qualification. There are in fact two distinct primacy claims that are rolled together in the kind of Pragmatism espoused by decision theorists. The first is that qualitative attitudes such as preference or comparative belief (attitudes of the form 'X is more credible/probable than Y') have primacy over quantitative ones such as degrees of desire or degrees of belief; the second that practical reason has primacy over theoretical reason. Both have methodological and conceptual dimensions.

The thesis of the methodological priority of qualitative attitudes over quantitative ones says that qualitative attitudes have methodological priority over the corresponding quantitative attitudes because our ability to attribute attitudes of the latter kind depend on our ability to determine attitudes of the former kind. That is, the qualitative attitudes provide the *evidence* for the quantitative ones. The corresponding thesis of conceptual priority says that they have conceptual priority because the rationality properties of the quantitative attitudes (such as degrees of belief being probabilities) derive from the rationality properties of the qualitative ones (such as transitivity of comparative belief). It is the fact

that the qualitative attitudes are rationally required to have certain relational properties that explains why the quantitative attitudes are rationally required to have corresponding numerical ones.

The priority claims for practical over theoretical reason run along similar lines. The methodological priority claim says that practical attitudes such as preference have methodological priority over theoretical ones such as belief because our ability to the determine the latter depends on our ability to determine the former. On the other hand the conceptual priority claims says that the laws of rational preference explain the laws of rational belief, both qualitatively and quantitatively.

With the exception of the last of them, I think that these priority claims are, with some qualifications, true. But I won't at this point argue directly for or against them. Rather I will focus on the motivation that they provide for decision-theoretic representation theorems and, conversely, on the question of what kind of support they derive from these theorems. First however the notion of preference requires further clarification.

4.2 Interpretations of Preference

Two broad classes of interpretations of the notion of preference can be found in the decision theoretic literature: those that define preference in terms of choice or behaviour and those that define it in terms of judgements or mental states. (Roughly the behavioural ones dominate in economics and the mentalistic ones in philosophy). More than one instance of each has been influential but I will simply spell out what seems to me the most viable versions of both.

Choice-theoretic: The basic thought underlying this class of interpretations is that preferences can be defined in terms of the choices or behaviour that they engender. Savage, for instance, regarded the claim that someone had a preference between two alternatives acts f and g as meaning that "if he were required to choose between f and g, no other acts being available, he would decide on f"(Savage, 1974/1954, p. 17). The view that Savage seems to be expressing has come to be called the Revealed Preference interpretation of preference. What it says, more formally, is that one alternative α is revealed-as-preferred to another β iff β is never chosen when both are available. Revealed Preference theorists sometimes speak as if they think that preferences are nothing but the choices that reveal them, but this talk is probably a result of a surfeit of positivistic enthusiasm for elimination of all reference to non-observable entities than a carefully thought-out position. For if preferences were nothing but choices then there would be nothing to reveal and certainly no sense in saying that preferences either explain or rationalise choices.

A more considered explication of the relationship between preference and choice captured in Savage's 'definition' is in terms of choice dispositions. On this account, a preference for α over β is a *disposition* to choose α rather than β when both are available. In contrast to the Revealed Preference account, on this view it is not analytic that β will never be chosen when α is available since dispositions have implicit normality conditions attached. Solubility-in-water is a matter of being disposed to dissolve when placed in water, but this disposition may not be revealed when the water is frozen, for instance. Similarly a preference for

α over β will not invariably eventuate in choice of α over β, for various other factors (error of judgement, unchecked emotions, etc.) might intervene in some contexts. Although preference is revealed in choice, not everything revealed in choice is preference.

The main advantage of the choice-theoretic construal of preference, and which explains its popularity in economics and other behavioural sciences, is that it requires the bare minimum of psychological assumptions in order to be applied. Preferences can be attributed to any entity which exhibits patterned choice, irrespective of its psychological constitution and complexity.[3] This gives decision theory enormous potential scope; indeed, there have been fruitful applications of it to animals, plants, machines and groups which exploit this flexibility of the preference concept. A second advantage of the choice-disposition approach is that it ties preference very closely to observable behaviour, thereby making it possible for rival models of preference to be tested empirically. This goes some way towards underpinning the methodological role that Pragmatism accords to preferences.

The main difficulty for choice-theoretic approaches, on the other hand, lies in accounting for the rationality properties that are usually attributed to preferences. Either it must be argued that these properties are embedded in the concept of choice itself or it must be granted that 'rational' preference is simply one subset of the kinds of preferences that a chooser might reveal. The behavioural turn in economics is testimony to the difficulty in making the former strategy work since there is evidence of patterned choice that lacks these properties. The latter strategy, on the other hand, leaves the representation theorems for behaviour without any normative role.

Judgementalism On judgementalist or mentalistic construals of preferences, they are a type of judgement or mental attitude.[4] As such they are the sorts of things that are susceptible to rationality conditions, an advantage of this interpretation over choice-theoretic ones. What kind of judgement are they? In fields such as welfare economics and social choice theory, they are typically taken to be judgements of personal wellbeing, so that to prefer one thing to another is to judge that it contributes more to one's wellbeing, all things considered. This restriction to what may be called self-interested preference, while perhaps appropriate in these fields, is not justified in general as one can clearly prefer things that are not in one's self-interest.

A better judgementalist interpretation of preference, defended by Dan Hausman (2011a)Hausman (2011b), is as an all-things-considered subjective comparative judgement.[5] The 'all-things-considered' part is crucial. In ordinary talk, we attribute preferences to describe someone's tastes, likings or favourings, which together with her moral beliefs, commitments, social norms and so on determine which action she chooses. On the all-things-considered notion of preference all of these constitute reasons for the agent to prefer one action over

[3] See Dennett (1971) for a revealing discussion of these explanatory virtues.

[4] In chapter 2, I argued that judgementalist and mentalistic interpretations of notions like probability are quite different. But I shall ignore these differences here.

[5] Hausman also takes them to be total rankings of alternatives, presumably because he wants to capture the notion of preference typically employed in economics. But the ability to form preferences over any two alternatives is not part of the concept of preference, so should not figure in its definition.

another and should be incorporated into their preferences. This brings preference closely into line with choice for if one's preferences incorporate all the reasons one has for favouring one action over another, then *ceteris paribus* one should choose it rather than the alternative. But it leaves open the possibility that extra-judgemental factors mediate the relationship between preference and choice.

I do not regard the best versions of these two classes of interpretation as rivals. Indeed, I favour a hybrid of them. On this hybrid account, a preference for α over β is best viewed as an all-things-considered comparative judgement that α is better than β that is instantiated in a disposition to choose the former over the latter when both are available (more generally, to have it be true that α rather than β). That preferences are judgements explains why they are subject to considerations of consistency. That they are also dispositions to choose both explains the connection between preference and choice and fixes the sense of betterness (namely as choice-worthiness) characteristic of preference judgements.

Both of these elements are essential. It is possible to make a choice-worthiness judgement but lack the disposition to choose and vice versa. But in neither case would it be appropriate to speak of preference. This is not to deny that some choice dispositions may never be revealed, since one can have preferences over alternatives between which one cannot choose (such as that one's grandchildren have happy lives). But this does not diminish the importance to the concept of preference of this relation between judgement and choice.

4.3 Representation Theorems

Representation theorems are the centrepiece of mathematical decision theory. But elegant mathematics aside, what do representation theorems achieve? In fact they play two different, but equally pivotal, roles in decision theory. The first is to provide a demonstration of how the values of the variables occurring within a decision theory (degrees of belief and desire, for instance) can be determined from information that can be gleaned from observation of behaviour. This is typically done in via the double-representation argument mentioned before. A first representation theorem establishes that observed choices meeting certain conditions determine attributions of preferences with certain properties (such as completeness and transitivity). A second representation theorem shows that preferences having these properties determine a measure of the agent's degrees of belief and desire.

The second role of a representation theorem for a decision theory, the one of most interest here, is to provide a justification of its normative claims regarding the properties of rational belief and desire and the relationship between them. It does so via an argument of the following kind.

- The axioms of preference either express rationality claims that all sensible people should accept or impose some structural conditions of little conceptual significance (but which are required for numerical representation).

- The representation theorem shows that satisfaction of these axioms by an agent's preferences implies the existence of a probability measure P and

utility measure U, respectively of her degrees of belief and her degrees of desire, that jointly represent these preferences.

- Therefore, since rationality requires her preferences to satisfy these axioms, it requires her to have degrees of belief and desire that are, respectively, probabilities and utilities.

Now an argument like this is not going to convince someone who has little sympathy for the Bayesian picture of rational agency even if they accept the premises. For they can simply deny that the numerical functions whose existence are established by the representation theorem are truly measures of the agent's degrees of belief and desire. They might accept that P measures a determinant of the agent's preferences in some formal sense, but deny that this determinant is the agent's real degrees of belief. They might, for instance, hold the view that rational preference is not sensitive to degrees of belief in the way that Bayesians claim and so conclude that P cannot be a measure of them. The Bayesian can retort that it is only the type of belief to which preference is sensitive that she cares to measure, but this is just a way of ending the discussion, not of convincing her opponent.

What this shows, I think, is that representation theorems can only play their justificatory role against the background of some shared assumptions about rational belief and desire and how these cohere with each other. They do not therefore give support to a decision theory by showing that its claims can be derived from first principles of preference whose motivation is entirely independent of that of the theory itself. Rather they serve to give it foundations by exhibiting the core qualitative principles upon which the theory depends, freed from the quantitative packaging in which it wrapped. A well constructed representation theorem will give a set of independent principles each expressing a feature of a widely shared conception of rationality, and derive implications from it of a much more precise nature. Its target therefore will be those who share the background assumptions that motivate the principles, but who do not necessarily accept the full corpus of Bayesian decision theory.

Even if we accept this non-foundationalist view of representation theorems, more domestic challenges remain. The main problem is that numerical representations of an agent's preferences are typically not unique. Not just in the sense that preferences do not determine the scaling of the numerical measures whose existence are established by the theorem, but in the more fundamental sense that they do not uniquely determine what form the numerical representation must take. For an agent's preferences may be also be numerically representable by other pairs of functions that, though not probabilities and utilities, can be combined in such a way as determine the preferences in question.[6] How are we to say which of these representations is the 'true' measure of the agent's degrees of belief and desire?

I don't think these questions admit of a single answer. Different representations will have different implications for the properties of the attitudes which they putatively measure and some of these implications may be more plausible than others. If this is so then ideally we should be able to pack these considerations into the conditions that underlie the representation so that the classes of representations it admits is more constrained. Equally there may be reasons for

[6] Or indeed by triples of functions, or any other number of them.

preferring one set of representations over another that do not derive from the attitudes we are trying to measure, but have to do with considerations of simplicity or technical convenience or continuity with other accepted theories. For example, the kinds of constraints that typical representation theorems impose on the representations of agent's degrees of belief do not uniquely determine that they should be represented by a probability function. Indeed, as we will see in the second part of the book, any real valued function that implies a certain kind ordering over prospects will do. The fact that a probability function assigns the value one to tautologous prospects and zero to contradictory ones is just an artifact of the scaling imposed by the numerical representation, a scaling chosen grounds of convenience. So it would be just plain silly to say that these theorems show that rationality requires degree of belief zero in contradictory prospects. What is not silly is to say is that rationality requires 'full' belief in the former and 'empty' belief in the latter, for these are properties that *are* picked out by their causal role in the determination of choice and which can be expressed as conditions on preference (see Bradley (2008)). All of this no doubt feels extremely abstract however, so let us turn to the examination of a particular instance of a representation theorem to get a better handle on these claims.

4.4 Savage's Representation Theorem

Although not the first representation theorem of its kind, that given by Leonard Savage in his book *Foundations of Statistics* (Savage (1974/1954)) is perhaps the most influential. In this section I will present his theorem, following his exposition quite closely, though with a few modifications for continuity with other sections.

Recall Savage's distinction between states, consequences and actions. States are complete descriptions of the features of the world that are causally independent of the agent's actions but which are relevant to determining their outcomes. Consequences, on the other hand, are the features of the world that matter to the decision maker, such as that she is in good health or wins first prize in a beauty contest or is allowed to sleep late on a Sunday morning. Actions are the link between the two, the means by which different consequences are brought about in different states of the world.

To reflect our earlier observation that the distinction between states and consequences is pragmatic rather than ontologicaI, I will treat both as elements of a single background set containing all the possible ways that the world might be—which I call prospects—rather than follow Savage in treating them as logically distinct types of object. Then the central elements of Savage's framework can be specified as follows:

1. $\Omega = \{A, B, C, ...\}$ is the set of prospects. Informally we can think of Ω as the set of all possibilities with respect to which an agent can have an attitude of belief or desire.

2. $\mathcal{C} \subseteq \Omega$ is the set of consequences. Informally we can think of $\mathcal{C}$ as a partition of Ω whose elements are maximally specific with regard to all that matters to the agent.

3. $\mathcal{S} = \{s_1, s_2, ...\}$ is the set of states of the world. Sets of states are called events. Informally we can think of $\mathcal{S}$ as a partition of Ω whose elements are maximally specific with regard to factors outside of the agent's control, but that are causally relevant to the determination of the consequences of acting.

4. $\mathcal{F} = \{f, g, h, ...\}$ is the set of actions.

5. $\succsim$ is the two-place 'at least as preferred as' relation on $\mathcal{F}$.

The goal of the representation theorem is to establish, from a set of conditions on the preference relation $\succsim$ on the set of acts the existence of a value function, V, on $\mathcal{F}$, which takes the form of an expected utility (i.e. $V(F) = \sum_{i=1}^{n} U(f(s_i)) \cdot P(s_i)$ for some real number function U on $\mathcal{C}$ and probability function P on $\wp(\mathcal{S})$) and which represents $\succsim$ in the sense that for all $f, g \in \mathcal{F}$:

$$V(f) \geq V(g) \Leftrightarrow f \succsim g$$

Savage proves the existence of an expected utility representation of preferences in two steps. First he postulates a set of axioms that are sufficient to establish the existence of a unique probability representation of the agent's beliefs. He then shows that probabilities can be used to construct a utility measure on consequences such that preferences amongst gambles cohere with their expected utilities, first on the assumption that the set of consequences is finite and then for the more general case of infinite consequences. Since the second step is essentially an application of Von Neumann and Morgenstern's theory, we will focus on the first and in particular on his derivation of a probability measure on events.

Recall that for Savage, actions are just functions from the set of states $\mathcal{S}$ into $\mathcal{C}$, the set of consequences. In fact Savage takes the preference relation to be defined over a very rich set of such acts, namely *all* functions from states to consequences. Because of its importance to his theorem, I have 'promoted' the definition of the domain of the preference relation to the status of an additional postulate.

P0 (*Rectangular field*)[7]: $\mathcal{F} = \mathcal{C}^{\mathcal{S}}$

Savage's first official postulate requires that the preference relation orders the sets of acts.

P1 (*Ordering*) $\succsim$ is (a) complete and (b) transitive.

For any consequence $F \in \mathcal{C}$, let act $\bar{f}$ be the corresponding constant act defined by, for all states s, $f(s) = F$. Given this definition it is straightforward to induce preferences over consequences from preferences over constant acts by stipulating that $F \succsim G$ iff $\bar{f} \succsim \bar{g}$. Such a stipulation in effect imposes the requirement that any feature of an act relevant to the preferences an agent has for it should be written into the consequences it determines.

Now P1 alone ensures the existence of a numerical representation V of $\succsim$ on $\mathcal{F}$ (when $\mathcal{S}$ is not countable P1 must be supplemented with a continuity condition, but let's just stick to the countable case). Hence a utility representation

[7] I take this term from Broome (1991).

U of $\succsim$ on $\mathcal{C}$ can be induced by setting, for all $F \in \mathcal{C}$, $U(F) = V(\bar{f})$. Note that it follows that V restricted to the constant acts trivially has the form of an expected utility since, in this case, the sum of the probabilities of the states determining the constant consequence of any such act is just one.

Savage's next step is to assume that the preference relation is separable across events i.e. that the desirability of a consequence of an act in one state of the world is ordinally independent of its consequences in other states. He does so by means of his famous Sure-thing Principle. Consider the acts displayed in the table below.

	Events	
Acts	E	E'
f	X	Y
g	X^*	Y

Intuitively, act f should be preferred to act g iff consequence X is preferred to consequence X^*. This is because f and g have the same consequence whenever E is not the case, and so should be evaluated solely in terms of their consequences when E is the case. Consequently any other actions f' and g' having the same consequence as f and g respectively whenever E is the case, and identical consequences when it is not, should be ranked in the same order as f and g. More formally:

P2 (*Sure-thing Principle*) Suppose that actions f, g, f' and g' are such that for all states $s \in E$, $f(s) = f'(s)$ and $g(s) = g'(s)$ while for all states $s \notin E$, $f(s) = g(s)$ and $f'(s) = g'(s)$. Then $f \succsim g$ iff $f' \succsim g'$

In view of P2 we can coherently define a conditional preference relation 'is not preferred to, given E', denoted $\succsim_E$, on the set of acts by, for all $f, g \in \mathcal{F}$:

$$f \succsim_E g \text{ iff } f' \succsim g'$$

where the acts f' and g' are as defined in P2. Conditional preference relations on consequences can then be induced in the same way as before. Given P1, it follows from this definition that each such conditional preference relation is complete and transitive.

The main role of P2 is to ensure that the consequences of an act in each state of the world can be evaluated separately. To see the implications of this, let the set of events $\{E_i\}$ be a partition of $\mathcal{S}$ and let f be an act that has the same consequence in every state in any E_i (hence $f(s) = f(s')$ iff s and s' belong the same element of the partition). Now in view of P1, P2 and the definition of conditional preference there exists numerical representations V_{E_i} of the preference relations $\succsim_{E_i}$ on $\mathcal{F}$ (again sticking to the countable case) and corresponding event-dependent utility measures U_{E_i} on $\mathcal{C}$ induced by setting, for all $F \in \mathcal{C}$, $U_{E_i}(F) = V_{E_i}(f)$. Note that this amounts to the choice of a representation satisfying $V_E(f') = V_E(f) \Leftrightarrow f'(s) = f(s)$ for all $s \in E$, i.e. to represent ordinally independent prospects as cardinally independent. Now P2, together with the choice, makes $V(f)$ a function of the $V_{E_i}(f)$; hence of the $U_{E_i}(F)$. In fact under some additional technical conditions it can be shown that it is possible to choose the V_{E_i} in such a way as to make $V = \sum_i V_{E_i}$ and hence such that $V = \sum_i U_{E_i}$, i.e. to represent the value of an act as the sum of

the event-dependent utilities of its consequences (see Krantz et al Krantz *et al.* (1971) for details).

What is now needed is a decomposition of the measure $U_{E_i}(F)$ into a probability for E_i and a utility for F. Then the additive representation will be revealed to be an expected utility. First an assumption is required to ensure the comparability of the event-dependent utilities. Let us call an event $E \in \Omega$ a null event iff $f \approx_E g$, for all $f, g \in \mathcal{F}$. Then Savage postulates:

P3 (*State Independence*) Let $B \in \Omega$ be non-null. Then if $f(s) = F$ and $f'(s) = G$ for every $s \in B$, then $f \succsim_B f' \Leftrightarrow F \succsim G$

The State Independence assumption ensures the *ordinal* uniformity of preferences from consequences across states, but is not strong enough to ensure the *cardinal* comparability of the state-dependent utilities. In particular, although it implies that, for any events E and E', $U_E(F) \geq U_E(G) \Leftrightarrow U_{E'}(F) \geq U_{E'}(G)$, it does not imply that $U_E(F) = U_{E'}(F)$. The next step is the crucial one for ensuring this as well as for obtaining a probability representation of the agent's attitudes to events. First Savage defines a 'more probable than' relation, $\trianglerighteq$, on the set of events. Consider the following pair of actions:

	Events	
Action	A	A'
f	X	Y

	Events	
Action	B	B'
g	X	Y

Actions f and g have the same two possible consequences, but f has the preferred consequence whenever A is the case and g has it whenever B is the case. Now suppose that consequence X is preferred to consequence Y. Then f should be preferred to g iff A is more probable than B because the action which yields the better consequence with the higher probability should be preferred to one which yields it with lower probability. More formally:

Qualitative probability: Suppose $A, B \in \Omega$. Then $A \trianglerighteq B$ iff $f \succsim g$ for all actions f and g and consequences X and Y such that:
(i) $f(s) = X$ for all $s \in A$, $f(s) = Y$ for all $s \notin A$,
(ii) $g(s) = X$ for all $s \in B$, $g(s) = Y$ for all $s \notin B$,
(iii) $X \succsim Y$

In effect the circumstances postulated by this definition provides a 'test' for when one event is more probable than another. Since it requires that $f \succsim g$ for any f and g meeting the conditions (i) - (iii), the existence of such a test does not itself guarantee that any pair of events can be compared in terms of their relative probability using this test. For this a further postulate is required.

P4 (*Probability Principle*) $\trianglerighteq$ is complete

In the presence of the other postulates, P4 ensures that preferences for actions depend on two factors only: preferences for consequences and the qualitative probability relation on events. It is not difficult to see that the definition of this latter relation implies that it is transitive. In fact, together with P4,

it also ensures that it is quasi-additive, i.e. that for all events C such that $A \cap C = \varnothing = B \cap C$:

$$A \trianglerighteq B \Leftrightarrow A \cup C \trianglerighteq B \cup C$$

Two further structural axioms are then required to ensure that the qualitative probability relation can be represented numerically.

P5 (*Non-Triviality*) There exists actions f and g such that $f \succ g$.

P6 (*Non-Atomicity*) Suppose $f \succ g$. Then for all $X \in \mathcal{F}$, there is a finite partition of S such that for all $s \in S$:
(i) ($f'(s) = X$ for all $s \in A$, $f'(s) = f(s)$ for all $s \notin A$) implies $f' \succ g$.
(ii) ($g'(s) = X$ for all $s \in B$, $g'(s) = g(s)$ for all $s \notin B$) implies $f \succ g'$.

P6 is quite powerful and implies that there are no consequences which are so good or bad that they swamp the improbability of any given event A. Nonetheless neither it nor P5 raises any pressing philosophical issues. And using them Savage proves:

Existence of Probability (Savage (1974/1954)) There exists a unique probability function P on $\wp(\mathcal{S})$ such that for all $A, B \in \wp(\mathcal{S})$:

$$P(A) \geq P(B) \Leftrightarrow A \trianglerighteq B$$

The rest of Savage's argument for existence of an expected utility representation of the preference relation applies von Neumann and Morgenstern's **?** representation theorem for preferences over lotteries. In essence what needs to be established is a correspondence between each act f and a lottery which yields each possible consequence C with probability, $P(\ f^{-1}(C))$, such that Savage's postulates for preferences over acts with a finite number of consequences imply that the induced preferences over the corresponding lotteries satisfy the Von Neumann and Morgenstern axioms. For then the value of each such act can be equated with that of the expected utility of the corresponding lottery. The proof is far from trivial however and I will not go any of the details here: see Savage (1974/1954) or Kreps (1988) very useful exposition.

4.5 Evaluation of Savage's axioms

In evaluating Savage's axioms it is useful to distinguish, in the manner of Suppes (2002), between those axioms expressing a requirement of rationality and those that play a technical or structural role in the proof of the representation theorem. In Suppes' view only P5 and P6 are structural axioms and the rest rationality conditions. In support of this classification, he notes that only these two conditions make existential demands and that neither is implied by the existence of an expected utility representation of preference. James Joyce (1999) adds P0, the rectangular field assumption, and P1a, the completeness assumption to the list of structural axioms. It is certainly clear that neither is a rationality condition. Furthermore both make existential claims—respectively about the richness of the action space and about the judgemental state of the agent—and neither is necessary for the existence of a numerical representation, though the latter is implied by standard expected utility representations.

	Ticket Numbers		
Actions	1	2 - 10	11 - 100
f	\$1000,000	\$1000,000	\$1000,000
g	\$0	\$5000,000	\$1000,000
f'	\$1000,000	\$1000,000	\$0
g'	\$0	\$5000,000	\$0

Table 4.1: Allais' Paradox

The only axioms that are unambiguously putative principles of rational preference are P1b, the transitivity condition, and P2, the Sure-thing Principle. P4, the Probability Principle, is plausibly a principle of rationality, but it is not really a fundamental principle of rational *preference*. Rather it is coherence constraint on the relation between belief and preferences. Finally P3—State Independence—is best regarded, not as a pure rationality claim, but as a constraint on the interpretation of consequences and states.

4.5.1 The Sure-thing Principle

The most discussed of Savage's axioms is undoubtedly the Sure-thing Principle. The main focus of attention in this regard has been apparent violation of the principle in the so-called Allais' paradox, a thought experiment proposed by Maurice Allais (1979). To illustrate it, consider two pairs of acts that are displayed in Table 4.1 which yield monetary outcomes conditional on the draw of a numbered ticket from a hat containing 100 different ones. Allais hypothesised that many people, if presented with a choice between actions f and g would choose f, but if presented with a choice between f' and g' would choose g'. Such a pattern of choice is, on the face of it, in violation of the Sure-thing Principle since the choice between each pair should be independent of the common consequences appearing in the third column of possible outcomes. Nonetheless Allais' conjecture has been confirmed in numerous choice experiments. Moreover many subjects are not inclined to revise their choices even after the conflict with the Sure-thing Principle is pointed out to them. So the 'refutation' seems to extend beyond the descriptive interpretation of the axiom to include its normative pretensions.

There are two lines of defense that are worth exploring. The first is to argue that the choice problem is under-described, especially with regard to the specification of the consequences. One common explanation for subjects' choices in these experiments is that they choose f over g because of the regret they would feel if they choose g and landed up with nothing (albeit quite unlikely), but g' over f' because in this case the fact that it is quite likely that they will not win anything whatever they choose diminishes the force of regret. If this explanation is correct then we should modify the representation of the choice problem faced by agents so that it incorporates regret as one possible outcome of making a choice. The same would hold for any other explanation of the observed pattern of preferences that refers to additional non-monetary outcomes of choices.[8]

[8] See Broome **?** for an extended defense of this kind.

The second line of defensive argument points to the gap between preference and choice. As we noted before, the specification of the choice set can influence the agent's attitudes. This is just such a case. In general the attitude we take to having or receiving a certain amount of money depends on our expectations. If we expect \$100, for instance, then \$10 is a disappointment. Now the expectation created by presenting the agent with two lotteries to choose from is quite different in the case where the choice is between lotteries f and g and the one in which the choice is between lotteries f' and g'. In the first case they are being placed in a situation in which they can expect to gain a considerable amount of wealth, while in the second they are not. In the first they can think of themselves as being given \$1000,000 and then having the opportunity to exchange it for lottery g. In the second case they can think of themselves as being handed some much lesser amount (say, whatever they would pay for lottery f') and then being given the opportunity to exchange it for lottery g'. Seen this way it is clear why landing up with nothing is far worse in the first case than in the second. It is because of what one has given up for it. In the first case landing up with nothing as a result of choosing g is equivalent to losing \$1000,000 relative to one's expectations, whereas in the second case it is equivalent to losing some much smaller amount.

Both of these defences are unattractive from the point of view of constructing a testable descriptive theory of decision making under uncertainty. The first approach makes it very hard to tell what choice situation the agents face, since the description of the outcomes of the options may contain subjective elements. The second approach makes it difficult to use choices in one situation as a guide to those that will be made in another, since all preferences are in principle choice-set relative. But from a normative point of view they go some way to supporting the claim that the Sure-thing Principle is a genuine requirement of rationality.

It is worth drawing attention to one further issue. As is evident from the informal presentation of the Sure-thing Principle, it is essentially a principle of weak dominance. That is to say that its intuitive appeal rests on the thought that since only the consequences of an action matter to its evaluation, if the consequences of one act are as least as good as those of another, and are better in at least one event, then this act is better overall. This application of consequentialist reasoning is not valid in general, however, for it will normally matter not just what consequences an action has, but how probable it makes them. Two actions could have identical consequences but if one of them brings about the better consequences with a higher probability than the other then it should be preferred to it. Such an eventuality is ruled out in Savage's framework because actions are construed as nothing but ordered sets of consequences, from which it follows that any two actions with the same consequences must have the same value (formally it is P0 which is doing the work here, by identifying the set of actions $\mathcal{F}$ with $\mathcal{C}^S$). So the Sure-thing Principle is compelling within his framework, but only because of the special form that actions take.

4.5.2 State Independence / Rectangular Field

It is not hard to produce apparent counterexamples to State Independence. Consider an act which has the constant consequence that I receive £100 and suppose I prefer it to an act with the constant consequence that I receive a

case of wine. Would I prefer receiving the £100 to the case of wine given any event? Surely not: in the event of high inflation for instance, I would prefer the case of wine. One could retort that receiving £100 is not a genuine consequence since its description fails to specify features relevant to its evaluation. Perhaps 'receiving £100 when inflation is low' might be closer to the mark. More generally, State Independence is bound to hold if we simply take consequences to be combinations of outcomes and the states in which they occur. But then the Rectangular Field assumption forces us to countenance actions which have such consequences in any state of the world, including those inconsistent with them. For example, it would require the existence of acts yielding £100 when inflation is low, in states of the world in which inflation is high. Such acts seem nonsensical and it is hard to see how anyone could express a reasonable preference regarding them.

An objection of this kind was famously made by Robert Aumann in a letter to Savage in 1971.[9] Savage's reply is interesting: he suggests that "a consequence is in the last analysis an experience"(Dreze, 1990, p. 79) and a state of the agent rather than of the world. The thought seems to be that experiences screen out the features of the world that cause them and hence have state-independent utilities. This is unpersuasive. however. On the whole I prefer that I be amused than saddened (or experiencing amusement to sadness), but I surely do not prefer it, given that a close friend has died. So even the desirability of experiences are contingent on the state of the world.

To the objection that his theory countenances nonsensical or impossible acts, Savage retorts that such acts "... serve something like construction lines in geometry" (Dreze, 1990, p. 79), and that they need not be available in order for one to say whether they would be attractive or not. But he seems to underappreciate the problem. Consider the decision whether or not to buy a life insurance policy that pays out some sum of money in the event of one's death. Now the pay-out is not a state-independent consequence in Savage's sense, for I am not indifferent between being paid while alive and being paid while dead. However the natural refinement of it gives us the consequence of 'pay-out and dead' which patently cannot be achieved in any state of the world in which I am alive. And even if I could summon a preference for being paid and alive, when alive, to being paid and dead, when alive, why does rationality require that my preference between these two consequences be the same for states in which I am dead? Perhaps, conditional on being dead, I am indifferent between the two.

In summary, State Independence is not plausibly a rationality constraint on preference. It is better to read it as a constraint on the specification of consequences, requiring in effect that they be sufficiently specific as to screen out, from an evaluative point of view, the state of the world. This difficulty with this interpretation is that it conflicts with the Rectangular Field assumption, since sufficiently specific consequences cannot occur in every possible state.

4.5.3 Probability Principle

Although P4 simply asserts the completeness of the 'more probable than' relation $\trianglerighteq$, it is the rationality claim underlying it that needs to be assessed. It is this: if two actions, such as f and g below, have the same two consequences then

[9] Printed, along with Savage's letter in reply, in Drèze (**?**, pp 76-81).

	red	black	yellow
L_1	\$100	\$0	\$0
L_2	\$0	\$100	\$0
L_3	\$100	\$0	\$100
L_4	\$0	\$100	\$100

Table 4.2: The Ellsberg Paradox

your preferences between them should depend only on the relative probability of the events determining the more preferred consequence. From this it follows, as illustrated below, that if $f \succsim g$, $X \succsim Y$, and $X^* \succsim Y^*$ then $f^* \succsim g^*$, as required by P4.

$$\text{If } \begin{array}{c|cc} f & A & A' \\ \hline & X & Y \end{array} > \begin{array}{c|cc} g & B & B' \\ \hline & X & Y \end{array} \text{ then } \begin{array}{c|cc} f^* & A & A' \\ \hline & X^* & Y^* \end{array} > \begin{array}{c|cc} f^* & B & B' \\ \hline & X^* & Y^* \end{array}$$

This rationale for P4 depends on the cardinal uniformity of utilities of consequences in different events because if event B makes both X more desirable than does event A then you could prefer g to f even if the probability of A was greater than that of B and $X \succ Y$. It also requires that the utilities of the consequences in one state be cardinally independent of the consequences of the act in other states. For example, assume that A and B are equiprobable, but that whether an act has consequence Y or Y^* in case of A' (and B') affects how much more desirable X is than Y, in case of A (and of B). Then you could prefer f to g but g' to f', because the desirability of X (respectively X^*) when it would have been the case that Y (respectively Y^*) if A had not been the case, is greater (less) than the desirability of X (X^*) when it would have been the case that Y (Y^*) if B had not been the case. Neither condition on the rationale for P4 can plausibly be said to be a purely formal one.

Most of the discussion of the Probability Principle has been directed elsewhere, at a thought experiment of Daniel Ellsberg Ellsberg (1961). In Ellsberg's experiment (see Table 4.2), an urn contains 90 balls, 30 of which are red, and the remaining 60 are black or yellow in an unknown proportion. Subjects are asked to choose between two bets. The first, L_1, pays off \$100 if, in a random draw from the urn, a red ball is drawn. The second, L_2, pays off \$100 if a black ball is drawn. Most subjects express a preference of L_1 over L_2. In a second choice problem, subjects are asked to choose between L_3 and L_4, which pay out \$100 in the events "red or black" and "black or yellow" respectively. Here, most subjects express a preference for L_4 over L_3.

It is evident that this pattern of preferences—the 'Ellsberg preferences' hereafter—violates the Sure-thing Principle as the 'yellow' column displays the same consequences for the two pairs of acts. But they are also inconsistent with the way in which Savage uses the Probability Principle to elicit subjective probabilities. For it follows from the definition of the qualitative probability relation that $L_1 \succ L_2$ iff the event 'red' is more probable than the event 'black' and that $L_4 \succ L_3$ iff the event 'black or yellow' is more probable than the event 'red or yellow'. But the laws of probability require that 'red' is more probable than 'black' iff for any event X disjoint with both, 'red or X' is more probable

than 'black or X'. Now strictly speaking the Probability Principle is not violated independently of Savage's other axioms. But the combination of the Probability Principle and the requirement that $A \trianglerighteq B$ iff $B' \trianglerighteq A'$ is. And it is hard to see what justification there is for the former that does not extend to the latter.

Most supporters of Savage has argued that the Ellsberg preferences are simply irrational and do not therefore constitute a refutation of his theory. Others have drawn the opposite conclusion: that the Ellsberg preferences are rational and that this fact shows that probabilistic degrees of belief are not rationally requirement in the kinds of situations of severe uncertainty exhibited in the Ellsberg setup. In chapter ?, I will argue for a third position: that the Ellsberg preferences are both rational and consistent with Savage's axioms. But some work is required to defend this position, so I will defer discussion of it.

4.6 Evaluation of Savage's argument

We have seen that Savage's representation theorem establishes that if an agent's preferences satisfy his postulates then she can be represented as a maximiser of subjective expected utility relative to a probability measure P on the set of events and a utility function U on the set of consequences. Our task now is to assess the significance, both methodological and normative, of this result. My focus here will be on the role that it plays within a pragmatist argument for Bayesian decision theory of the kind sketched at the beginning of the chapter. But first let me first consider the plausibility of a behaviourist interpretation of it.

To provide support for the methodological claims of Behaviourism, one might read Savage's postulates as conditions on choice (in concert with a Revealed Preference interpretation) and hence his theorem as showing that if they are satisfied by observed choices then quantitative degrees of belief and desire can be attributed to the chooser. But this idea faces a fundamental problem. Savage's theorem starts with preferences over acts, construed as functions from states to consequences. But *although we can observe choices amongst acts, we cannot observe what the acts are that the agent is choosing between.* This is because we cannot tell from an agent's choices how they conceive of the objects of choice; in particular what consequences they believe to follow from this choice in each possible state of the world. Choice reveals preferences only with a framework of common representation of the objects of choice. It would be natural to achieve this by verbal descriptions of the objects of choice. But if recourse must be had to verbal communication then why not simply ask the subjects what they prefer and dispense with the pretence of purely behavioural evidence?

Behaviourism is perhaps something of a straw figure here (even though its influence in decision theory is considerable) since it doubtful that Savage was committed to it, so let us turn to an assessment of his theorem within a pragmatist framework. There are two claims that need to be assessed: the methodological claim that Savage's theorem establishes sufficient conditions for attribution of degrees of belief and desire to agents on the basis of their preferences, and the normative claim that his theorem establishes that rationality requires agents to have probabilistic degrees of belief and to maximise expected utility relative to them. Both claims depend, first, on the status (empirical or normative) of his axioms of preference and, second, on the import of the demonstration of the ex-

istence of a particular kind of numerical representation of preferences satisfying them.

Our brief discussion of the first issue did not entirely settle the question of whether all his axioms are either genuine rationality conditions or else 'harmless' structural ones. It is clear that the combination of State Independence and the Rectangular Field assumption is problematic from both an empirical and a normative point of view and it would be better for the pragmatist argument if it could be dispensed with. Perhaps the same is true for the Sure-thing Principle, at least empirically, but I shall set aside further consideration of it until later in the book. More pressing is the status of the Probability Principle, for it is clearly not purely a principle of rational preference. Rather it is rationality condition on the relationship between preferences and qualitative beliefs. Furthermore it is not a condition that anyone is likely to accept unless they independently adhere to the view about the role of belief in preference formation that it expresses. Given the central role that it plays in the derivation of degrees of belief from preferences, this somewhat vitiates the claim that Savage's representation theorem supports the conceptual priority of practical reason, since it undermines the claim that the properties of rational belief can be derived from independent properties of rational preference. (This objection does not of course apply to the other priority thesis).

Although much of the literature on Savage's theory is focused on the status of axioms, a more fundamental problem derives from the restrictive nature of Savage's framework. Recall that states of the world must be probabilistically independent of the acts over which preferences are defined. So degrees of belief can be inferred from preferences only for such states. But there will be many features of the world that are not independent of the agent's actions that she nonetheless has beliefs about; for instance, whether she will perform any particular action or not! Since such beliefs fall outside the scope of Savage's theorem, he cannot be said to have established either that all degrees of belief are measurable from preferences or that all partial belief must be probabilistic. A similar point can be made about the restriction of utilities to maximally specific consequences. In a nutshell Savage's method doesn't yield measures of degrees of belief and desire for all prospects (the background set Ω), but only of those belonging in either $\mathcal{S}$ or $\mathcal{C}$.

There is a second, more subtle, problem. Although Savage shows that preferences satisfying his postulates have an expected utility representation unique up to a choice of scale for the utility function, he does not show that such a form of representation is unique. Indeed it quite clearly isn't. For any agent who maximises expected utility relative to a probability measure P and utility U also maximises a function V' defined by:

$$V'(f) = \sum_{i=1}^{n} \alpha_i U(f(s_i)) \cdot P'(s_i)$$

where $\alpha_i = nP(s_i)$ and $P'(s_i) = \frac{1}{n}$. So the agent can be represented as if she assigns equal probability to each state of the world, but has utilities for consequences that vary with the state in which they occur. The choice of a cardinally state-independent representation is an entirely arbitrary one, not required by Savage's axioms (recall that State Independence is a purely ordinal requirement). This being so we have no axiomatic basis for saying that the agent's preferences determine that her degrees of belief are measured by P rather than

P'.

The problem can be expressed in a slightly different way. The Probability Principle allows a qualitative probability ordering over events to be constructed from preferences over acts. But why should we take the constructed ordering to represent the agent's relational beliefs? The most plausible answer is that the constructed ordering plays just the role in the determination of the agent's preferences that one would expect of them. But as I showed before, this argument presupposes that an agent's preferences for the consequences of her actions should be cardinally independent of both the state in which they occurs and of the counterfacts, for if they were not then her beliefs would not combine with her desires in the manner required by the Probability Principle. Neither presupposition is guaranteed by Savage's postulates. So even if we regard the Probability Principle as a rationality constraint we cannot infer that the qualitative probability relation constructed from the preference relation in accordance with it correctly represents the agent's relational beliefs and so we cannot conclude that she maximises expected utility relative to them. In summary, an agent who satisfies Savage's postulates maximises expected utility relative to some probability P, but not necessarily relative to her actual degrees of belief.

Savage's representation theorem does not, it seems, deliver the goods. To plausibly interpret his theorem in the manner required by the pragmatist argument for Bayesian decision theory, states and consequences must be specified in a way which ensures the cardinal independence of preferences for consequences from both states and the counterfacts. But this then considerably restricts the scope of any conclusions that can be drawn from his theorem, for many prospects with respect to which we have both beliefs and desires will not meet these conditions. To give foundations to Bayesian decision theory, we will have to do better. In the next part of the book, I hope to do so.

Bibliography

Allais, Maurice. 1979. The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School. *In:* Allais, Maurice, & Hagen, Ole (eds), *Expected Utility Theory and the Allais Paradox: Contemporary Discussions of Decisions under Uncertainty with Allais' Rejoinder*.

Binmore, Ken. 2008. *Rational decisions*. Princeton University Press.

Bradley, Richard. 2008. Comparing Evaluations. *Proceedings of the Aristotelian Society*, **108**(1), 85–100.

Bradley, Richard, & Stefansson, H Orii. 2016. Desire, expectation, and invariance. *Mind*.

Broome, John. 1991. *Weighing Goods*. Basil Blackwell.

Broome, John. 1999. Can a Humean be Moderate? *In: Ethics Out of Economics*. Cambridge University Press.

Dennett, Daniel C. 1971. Intentional systems. *The Journal of Philosophy*, 87–106.

Dreze, Jacques. 1990. *Essays on economic decisions under uncertainty*. CUP Archive.

Ellsberg, Daniel. 1961. Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, **75**(4), 643–669.

Hacking, Ian. 2006. *The emergence of probability: a philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press.

Hausman, Daniel M. 2011a. Mistakes about preferences in the social sciences. *Philosophy of the social sciences*, **41**(1), 3–25.

Hausman, Daniel M. 2011b. *Preference, value, choice, and welfare*. Cambridge University Press.

Jeffrey, Richard. 1990/1983. *The Logic of Decision*. The University of Chicago Press (revised edition).

Joyce, James. 1999. *The Foundations of Causal Decision Theory*. Cambridge University Press.

Karni, Edi. 1985. *Decision Making under Uncertainty: The Case of State-Dependent Preferences*. Harvard University Press.

Krantz, David, Luce, R. Duncan, Suppes, Patrick, & Tversky, Amos. 1971. *Foundations of Measurement. Volume 1. Additive and Polynomial Representations*. Academic Press.

Kreps, David. 1988. *Notes on the Theory of Choice*. Westview press.

Lewis, David. 1981. Causal Decision Theory. *Australasian Journal of Philosophy*, **59**(1), 5–30.

Lewis, David. 1988. Desire as Belief. *Mind*, **97**(387), 323–32.

Lewis, David. 1996. Desire as Belief II. *Mind*, **105**(418), 303–313.

Meacham, Christopher J. G., & Weisberg, Jonathan. 2011. Representation Theorems and the Foundations of Decision Theory. *Australasian Journal of Philosophy*, **89**(4), 641–663.

Parfit, Derek. 2013. *On what matters: volume one*. Vol. 1. Oxford University Press.

Ramsey, Frank P. 1990. Truth and probability (1926). *Pages 156–198 of: FP Ramsey: Philosophical Papers*. Cambridge University Press.

Ramsey, Frank P. 1990/1926. Truth and Probability. *In:* Mellor, D. H. (ed), *Philosophical Papers*. Cambridge University Press.

Savage, Leonard. 1974/1954. *The Foundations of Statistics*. Dover Publication (revised edition).

Simon, Herbert A. 1957. Models of man; social and rational.

Simon, Herbert A. 1986. Rationality in psychology and economics. *Journal of Business*, S209–S224.

Simon, Herbert A. 1990. Bounded rationality. *Pages 15–18 of: Utility and probability*. Springer.

Stalnaker, Robert. 1981. Letter to David Lewis. *In:* Harper, William L., Stalnaker, Robert, & Pearce, Glenn (eds), *Ifs: Conditionals, Belief, Decision, Chance, and Time*. D. Reidel Publishing Company.

Starmer, Chris. 2000. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of economic literature*, 332–382.

Suppes, Patrick. 2002. *Representation and invariance of scientific structures*. CSLI publications Stanford.

von Neumann, John, & Morgenstern, Oskar. 2007/1944. *Games and Economic Behavior*. Princeton University Press.

Walker, Oliver, Dietz, Simon, *et al.* . 2011. A representation result for choice under conscious unawareness. *Centre for Climate Change Economics and Policy Working Paper 68, Munich Re Programme Technical Paper*, **10**.

Weirich, Paul. 2004. *Realistic decision theory: Rules for nonideal agents in nonideal circumstances*. Oxford University Press.

Zynda, Lyle. 2000. Representation theorems and realism about degrees of belief. *Philosophy of Science*, 45–69.

26502 words