Introduction

Imagine yourself in the following situation. There are two boxes before you: one transparent and one opaque. You can see that there is $1,000 in the transparent box, and you know that there is either $1,000,000 or nothing in the opaque box. You must choose between the following two acts: take the contents only of the opaque box or take the contents of both boxes. Furthermore, there is a being in whose predictive powers you have enormous confidence, and you know that he has already determined the contents of the opaque box according to the following rules: If he predicted that you would take the contents only of the opaque box, he put the $1,000,000 in the opaque box, and if he predicted that you would take the contents of both boxes, he put nothing in the opaque box. What would you do?

A paradox, known as 'Newcomb's paradox', seems to arise. Robert Nozick (1969) discusses the paradox in detail. It appears that two principles of decision – both of which are well-respected and intuitively attractive – prescribe different courses of action in the decision situation described above. Consider this version of the *principle of dominance*: If (i) you must perform either act A or act B, (ii) which act you perform does not causally affect which of two states of affairs, S and $-S$, obtains, and (iii) no matter which of S and $-S$ obtains, you are better off doing A than doing B, then do A. In the decision situation described above, (i) you must either take the contents only of the opaque box or take the contents of both boxes, (ii) which act you perform does not affect whether or not the $1,000,000 is in the opaque box, and (iii) whether the $1,000,000 is in the opaque box or not, you get $1,000 more by taking the contents of both boxes than you get by taking the contents only of the opaque box. So the principle of dominance recommends taking the contents of both boxes.

Now consider this rough statement of the *principle of maximizing conditional expected utility* (hereafter, *PMCEU*): perform the act that makes the most desirable outcomes the most probable. In the decision situation described above, if you take the contents only of the opaque box, then,

since the predictor is so accurate, the predictor probably predicted you would do that, in which case he would have put the $1,000,000 in the opaque box and you would walk away with $1,000,000. If you take the contents of both boxes, then, since the predictor is so accurate, the predictor probably predicted you would do *that*, in which case he would have left the opaque box empty and you would walk away with only $1,000. *PMCEU* seems to prescribe taking only the contents of the opaque box.

Several philosophers who believe, as I do, that the correct act is to take the contents of both boxes believe, as I do not, that *PMCEU* should be given up. Recently a number of other *prima facie* counterexamples to *PMCEU*, all inspired by Newcomb's paradox, have been constructed. These involve decision situations that are less fantastic than that of Newcomb's paradox and in which almost everyone would agree that the correct act is the counterpart of the two-box act in Newcomb's paradox. Central to the alleged counterexamples is the observation that *PMCEU* seems not to be sensitive to causal beliefs of a certain kind that an agent might have, which suggests that some causal notions need to be introduced into the calculation of expected utility. Thus, various "causal decision theories" have recently emerged as rivals to *PMCEU*.

The first three chapters of this book provide an exposition of the more general philosophical ideas and theories in terms of which the controversy surrounding Newcomb's paradox, *PMCEU* and causal decision theory is philosophically significant. Chapter 1 is an introduction, from the decision-theoretic point of view, to the philosophical view known as 'Bayesianism'. In Chapter 2, the philosophical significance and empirical adequacy of Bayesian decision theory are explored. And Chapter 3 presents, in more detail, several versions of traditional ('noncausal') Bayesian decision theory, of which *PMCEU* is one.

In Chapter 4, I present a number of *prima facie* counterexamples to *PMCEU* of the kind inspired by Newcomb's paradox, and I try to clarify their causal structure. Chapter 5 deals with some of the new causal decision theories. In it, I argue that a successful *PMCEU* approach to the problem would have important advantages over the causal approach. And in Chapters 6 and 7, I argue that *PMCEU* really gives the correct prescriptions in the decision situations of the alleged counterexamples and in general – indeed, the same prescriptions given by the principle of dominance and by causal decision theory. Chapter 8 deals with Newcomb's paradox itself in detail.

1

Bayesianism

Bayesianism is usually characterized as the philosophical view that, for many philosophically important purposes, probability can usefully be interpreted subjectively, as an individual's "rational degree of belief," and that the rational way to assimilate new information into one's structure of beliefs is by a process called 'conditionalization'. The subjective interpretation of probability is connected, however, in very important ways with a mathematically precise and intuitively plausible theory of rational decision, called the 'subjective expected utility maximization theory'. Because of this connection, and the nature of it, Bayesianism can alternatively be characterized as the view that (i) rational decision and rational preference go by subjective expected utility, (ii) subjective probabilities (and numerical subjective utilities) are more or less theoretical entities that "lie behind," explain and are given partial empirical interpretation by, an individual's choices and preferences and (iii) learning goes by conditionalization. In this chapter and the next two, I will describe these three aspects of Bayesianism, discuss their plausibility and indicate various ways in which they are philosophically significant. The subsequent chapters will deal with a formidable challenge to this potentially very powerful philosophical theory.

Subjective expected utility

Deliberation is the process of *envisaging* the possible consequences of pursuing various possible courses of action and *evaluating* the merits of the possible courses of action in terms of their possible consequences. Roughly, the Bayesian model says that a course of action has merit to the extent that it makes good consequences probable and that a rational person pursues a course of action that makes the best consequences the most probable, where the goodnesses and probabilities of the consequences are the agent's subjective assessments thereof: how true, reasonable or otherwise objectively or morally sound these assessments are is regarded as a separate question.

This last point is quite important; it indicates an essential feature of Bayesianism and Bayesian decision theory which is worth fully noting at the outset. It is implicit in the theory that whether or not a given course of action in a given decision making situation is rational is not an absolute kind of thing: a course of action is rational only relative to a possessed body of information (beliefs and desires) in terms of which the merits of the available courses of action can be rationally evaluated. Properly conceived, therefore, decision making involves two processes: (i) obtaining a body of relevant information (the process of information-acquisition) and (ii) evaluating the available courses of action in terms of the information at hand (the process of deliberation, or of information-use). We may say that an action is *rational* to the extent to which process (ii) is successfully carried out. And we may say that an action is *prudential* (or rational *and* well-informed) to the extent to which *both* processes are successfully carried out and, therefore, to the extent to which the action is truly, objectively in the agent's best interest to perform. Bayesian decision theory is primarily concerned with part (ii) of the decision making process. (Since the activity of information-gathering itself involves decisions, however, it is not surprising that the theory has also been applied (e.g., by Adams & Rosenkrantz 1980) to part (i) of the decision making process.) It is a theory about how one's actions, preferences, values and beliefs must be related *to each other* – not how they should be related to the objective world – for them to be rationally so related. Thus, Bayesian decision theory is as applicable to the deliberation of the ignorant and inexperienced as it is to that of the knowledgeable expert; and it is as applicable to the deliberation of a monster as it is to that of a saint.

Before considering a precise general statement of the theory, consider this concrete example, adapted from Richard Jeffrey's *The Logic of Decision* (1965*b*). You are to be the dinner guest of some acquaintances tonight, and you are to provide the wine. You remember that they plan to serve either chicken or beef, but you have forgotten which, although you know that they typically serve chicken. You have a bottle of white and a bottle of red, no telephone and can bring only one bottle, as you are going by bicycle. Associated with this decision situation are three matrices: the *outcome* (or *consequence*) *matrix*, the *desirability matrix* and the *probability matrix*. The consequence matrix for this decision problem may be:

	Chicken	Beef
White	White wine with chicken	White wine with beef
Red	Red wine with chicken	Red wine with beef

It indicates the possible outcomes, or consequences, of each possible course of action. The two row-headings indicate the available *acts,* the column-headings the possible *states* and the entries the *outcomes* that result from performing a given act under a given state. The desirability matrix, plausibly

	Chicken	Beef
White	10	−10
Red	0	10

indicates the numerical *desirabilities,* or *utilities,* that correspond to the entries in the outcome matrix. These desirabilities are subjective in the sense that they represent the agent's assessments of the desirabilities of the outcomes. Finally, the probability matrix, say

	Chicken	Beef
White	0.6	0.4
Red	0.6	0.4

indicates subjective assessments of the probabilities of the outcomes, assuming that the act which the entry-row represents is performed. Note that the entries in each row add to 1.

The subjective expected utility of the acts is calculated as follows. First, multiply corresponding entries of the desirability and probability matrices. The result in this case, dropping the column headings, is:

White	6 −4
Red	0 4

Then add the entries in each row to get:

White: 2
Red: 4.

Thus, the subjective expected utility of bringing the bottle of white is 2; that of bringing the red is 4. In symbols, *SEU*(White) = 2 and *SEU*(Red) = 4. The subjective expected utility maximization theory (*SEU* theory, for short) recommends bringing the red wine.

Note that in the above example, the two rows in the probability matrix are identical. This is only a special case. For suppose that even though

your hosts prefer chicken and typically serve chicken to guests, they may be influenced by (among other things) your choice of wine. In this case, the probability matrix might be:

	Chicken	Beef
White	0.9	0.1
Red	0.3	0.7

Relative to this probability matrix and the old desirability matrix, *SEU* (White) = 8 and *SEU*(Red) = 7 so that *SEU* theory recommends bringing the white wine. In this case, because of your hosts' cooperation, the *SEU*s of both acts are higher than those in the previous case.

Thus, to apply *SEU* theory to a decision problem, the decision situation must first be represented in terms of outcome, desirability and probability matrices. The possible outcomes of an act A can be denoted by 'O_{Ai}'; the desirability of O_{Ai} can be denoted by 'd_{Ai}'; and the probability of getting the outcome O_{Ai} when act A is performed can be denoted by 'p_{Ai}'. Note that the outcomes have act-subscripts. This is reasonable because the ultimate carriers of desirability are not just the things you get, independently of the act: they involve also *how* you get them (Adams & Rosenkrantz 1980). This was suppressed in the example given above; but, if the example were worked out in more detail, one might wish to distinguish between the outcome of bringing white wine and drinking it with chicken and the outcome of bringing red wine and drinking (perhaps your hosts') white wine with chicken. Of course, the ultimate carriers of desirability could alternatively be symbolized by expressions like 'O_i & A', which symbolizes the "act-specific" outcome of doing A and getting the non-act-specific O_i.

An alternative way of denoting probabilities and desirabilities is by using function, or assignment, symbols: say 'P' and 'D', respectively. Thus, instead of writing 'd_{Ai}', we could write '$D(O_{Ai})$' (or '$D(O_i$ & $A)$'). As to the p_{Ai}s, various suggestions have been made as to what precisely they should be the probabilities of. On one suggestion, p_{Ai} is the probability of O_{Ai} conditional on A, i.e., $P(O_{Ai} \mid A) = P(O_{Ai}$ & $A)/P(A)$ (or the probability of O_i conditional on A, $P(O_i \mid A)$); on another, it is the unconditional probability of the *state* under which performing A results in the outcome O_{Ai} (or O_i); and on a third, it is the probability of the counterfactual conditional 'If A were performed, then O_{Ai} (or O_i) would result'. In Chapters 3 and 5, we will look at a number of suggestions. Here, I just want to present the basic idea which is common to all the ways in which Bayesian decision theory has been developed.

Given the entries of the desirability and probability matrices, we can calculate the subjective expected utility of an act A as follows:

$$SEU(A) = \sum_I p_{Ai} d_{Ai}.$$

The *SEU* theory asserts that rational preference goes by *SEU* – i.e., that an act A is rationally preferred to an act B if, and only if, $SEU(A) > SEU(B)$ – and that a rational person chooses a course of action that has the greatest *SEU*. The *SEU* theory can be interpreted normatively, descriptively, or both. In the next chapter, the normative and descriptive adequacy and significance of the theory will be discussed.

It should be borne in mind that the present statement of *SEU* theory is rough and sidesteps some important issues, such as the nature of the distinction between states and outcomes and the possibility that an act and state together do not determine a unique outcome but rather may result in different outcomes with different probabilities. Such issues as these will be dealt with in Chapter 3, which discusses various detailed ways in which Bayesian decision theory has been developed.

Foundations of subjective probability

Subjective probabilities, or degrees of belief, are, on the decision-theoretic analysis, more or less theoretical entities which, together with subjective desirabilities, "lie behind" and explain the more or less observable phenomena of preference and choice. In this respect, subjective probability theory stands in the same relation to preference and choice as, for example, the kinetic theory of gases (about the behavior of the molecules) stands to the behavior of gases. The philosophical foundations of the view that belief is just that which is so related to preference and choice are to be found in the dispositional theory of belief, which will be discussed in the next chapter. But something more is needed for foundations of a theory of *degrees* of belief, when those degrees are, on the theory, *probabilities.* That something more is provided by various "representation theorems" (Savage 1954, Bolker 1967, Domotor 1978, Jeffrey 1978). These theorems relate preference data to a pair of functions: a probability function, which is, plausibly, the relevant agent's subjective probability assignment, and another, which is, plausibly, the agent's subjective desirability assignment. I will give the general idea of the theorems without fully stating them or discussing them in detail.

The theorems are all to the effect that if a preference relation (i.e., 'is preferred to') satisfies certain axioms (which in general are intuitively

plausible), then there exists a probability function P and another function D such that, when SEU is calculated in terms of them, an item X will be preferred to an item Y if, and only if, $SEU(X) > SEU(Y)$. The nature of the entities over which the preference relation is defined depends on the detailed way in which SEU theory is developed. On Ramsey's (1926) and Savage's (1954) theories (discussed in Chapter 3), they are gambles, i.e., options of the form: you get consequence O_1 if proposition p is true and consequence O_2 otherwise. And on Jeffrey's (1965*b*) theory (also discussed in Chapter 3), they are propositions. It should also be noted that the preference data do not determine *unique* functions P and D. On Ramsey's and Savage's theories, P is uniquely determined, but D is unique only up to linear transformations, i.e., if D and D' are both derivable from the preference data, then there exist real numbers a and b such that for any item X over which the preference relation is defined, $D'(X) = aD(X) + b$. On Jeffrey's theory, neither P nor D is uniquely determined – a family of pairs of functions is determined.

Also, it might be useful to point out the basic insight as to how probability data can be derived just from preference data. If an agent prefers O_1 to O_2 and he prefers the gamble in which he gets O_1 if p, O_2 otherwise, to the gamble in which he gets O_1 if q, O_2 otherwise, then this must be because he thinks that p is more probable than q; clearly one prefers to stake one's chances of getting the more desirable outcome on the most probable proposition. (Note that if preference is defined only on gambles, we can still think of an outcome O as the "gamble" in which you get O if $p \vee -p$ is true and O' otherwise, or as the "gamble" in which you get O if p, O otherwise.)

There are two kinds of axioms that the representation theorems assume a preference relation to satisfy, sometimes called 'the *necessary* axioms' and 'the *nonnecessary* axioms'. The necessary axioms are consequences of the conclusion of the theorems: conditions that must be satisfied for the conclusion of the theorems to be true. It is obvious, for example, that the conclusion of the theorems implies that the preference relation must be *transitive*. Letting '$\succ$' denote the relation of preference (i.e., '$X \succ Y$' means 'X is preferred to Y'), the conclusion of the theorems is that there exists a probability function P and a (desirability) function D such that, when SEU is calculated in terms of these assignments, then $X \succ Y$ if, and only if, $SEU(X) > SEU(Y)$, for all X and Y. Thus, since $>$ is transitive, $\succ$ must be as well. Also, since $>$ is *trichotomous*, it is obvious that $\succ$ must be as well, i.e., for every X and Y, either $X \succ Y$ or $Y \succ X$ or $X \sim Y$ (where '$X \sim Y$' indicates indifference, i.e., not $X \succ Y$ and not $Y \succ X$, and the 'or's are "exclusive").

Table 1

	S	$-S$
A_1	O_1	O_3
A_2	O_2	O_3
B_1	O_1	O_4
B_2	O_2	O_4

Another important and, as we shall see later, somewhat controversial necessary postulate is what Savage calls 'the sure-thing principle': If two acts have the same outcome in a particular state of nature, then which act one prefers should be independent of what that outcome is. More precisely, if, for example, the relevant outcome matrix is as in Table 1, then, according to the sure-thing principle, $A_1 \succ A_2$ if, and only if, $B_1 \succ B_2$. Assuming that the decision problem is formulated in such a way that the acts do not affect the probabilities of the states (Chapter 3 indicates a way in which this can be done), it is easy to see that the conclusion of the representation theorems implies the sure-thing principle.

Most empirical research on the descriptive adequacy of Bayesian decision theory consists of experimentally testing subjects' conformity to one or more of the necessary postulates. In the next chapter, some of this research will be described.

The nonnecessary axioms of a representation theorem are of a technical nature and assert that the set of acts, states and outcomes satisfy certain formal, structural conditions, not all of which involve the preference relation. The nonnecessary axioms are sometimes called 'structural axioms'. I shall not discuss these until, in Chapter 3, we look at some of the detailed ways in which *SEU* theory has been developed. Meanwhile, two examples might help to clarify their nature. Bolker (1967) assumes that the set of propositions involved in Jeffrey's decision model is an atomless Boolean algebra (see Appendix l), the main effect of which is that for any proposition X, there is a nonequivalent proposition Y which implies X. And Savage (1954) assumes that for every outcome, there is an act which invariably results in that outcome.

The main significance of the representation theorems for our purposes is that, by indicating how subjective probabilities can be measured, they, together with the *SEU* theory and a dispositional theory of belief, provide foundations for the theory of subjective probability. Assuming a dispositional theory of belief and the correctness of the *SEU* theory, the representation theorems give partial empirical interpretation to subjective

probabilities, since preference and choice are observable phenomena. In the next chapter, I will delineate this idea in more detail, contrast it with other approaches and indicate the potential power and importance of a well-founded theory of subjective probability.

Learning

Bayesianism has a static part and a dynamic part. The static part asserts that rational degrees of belief can be represented by a probability assignment over propositions (or events, gambles, etc.). One way of justifying this static part is along the lines sketched just above, relying on the intuitive attractiveness of the preference axioms and on a dispositional theory of belief; other suggested justifications will be considered later. The dynamic part of the theory asserts that rational change of (degrees of) belief takes place in a certain way.

Now, just as Bayesian *decision* theory tells you what course of action it is rational to pursue *relative* to your beliefs and desires, irrespective of how factually or morally justified they may be, so Bayesian *learning* theory tells you what new degree of belief assignment it is rational to adopt when new evidence comes in *relative* to what your prior degrees of belief are. Just as decision making involves both past information-acquisition and present deliberation, so changing your degrees of belief involves both (i) having already adopted a prior degree of belief assignment and (ii) changing it to accommodate the new evidence. The adoption of a particular posterior assignment which accommodates new evidence may be said to be a *rational* move to the extent to which process (ii) is successfully (i.e., favorably, validly, correctly) carried out; the move may be said to be *well-grounded* to the extent to which *both* processes are successfully (favorably, validly, correctly) carried out and, thus, to the extent to which the posterior assignment accommodates not only the recently acquired evidence, but also previous experiences of the agent. Bayesian learning theory is concerned with part (ii) of the process of belief change. It is a theory about how the new assignment must be related *to the old one* – and not how it must be related to the objective world – for it to be rationally so related just in virtue of the acquisition of the new evidence. Thus, the learning theory is as applicable to the learning undergone by the ignorant and inexperienced as it is to that undergone by the knowledgeable expert.

The dynamic part of Bayesian theory asserts that *rational change of belief goes by conditionalization*: i.e., that if one learns that some proposition E is