B.1 The general idea of decision theory

How do rational agents act?

One way to answer the question begins at the end, with the well-known principle that rational agents should act by following the rule of maximizing expected utility (MEU). That is, if $c_1, \ldots c_n$ are the possible consequences of an agent's action, he makes use of a *utility function* $\mathcal{U}$ which maps these consequences to real numbers, a *probability function* Pr which gives the probability of each consequence actually happening, and he chooses whichever action A maximizes

$$\sum_i \Pr(c_i|A)\mathcal{U}(c_i). \tag{B.1}$$

The MEU rule is often presented as if Pr and $\mathcal{U}$ are just given to us (as if the utility of a consequence is its cash value, say, and the probability of that consequence is something given to us by the laws of physics). From this perspective, agents look up the probabilities and utilities, and then (if they are rational) they choose whichever action is best according to the MEU rule. Probabilities and utilities are inputs, preferences are outputs.

From this perspective, though, it is quite mysterious *why* maximizing expected utility is the right thing to do. Why not maximize the expected value of $(\text{utility})^2$, or log(utility), for instance? Equally, where do these 'utilities' and 'probabilities' come from?

In the Dutch book argument sketched in Chapter 4, we saw an alternative strategy. In a Dutch book argument, we do take utility as an input: in fact, we take the utility of a reward to be its value in dollars. But we do not take probability as an independent input: instead, we define the probability of an event E as the maximum price, in dollars, we would pay for a bet which returns one dollar if and only if E obtains. We then argue that, unless those probabilities obey the axioms of probability calculus, the agent is committed to accepting bets which cause him to lose money.

The Dutch book method for defining probability makes MEU true by definition, at least in the case of these simple bets: the expected utility of a one-dollar bet on E is the probability of E, which by definition is the cash value of that bet. And if (as the Dutch book argument tacitly assumes) the cash value of a combined bet is the sum of the cash values of its components, this can be extended to more complicated bets (ones which return different amounts on different outcomes).

It might appear, then, that the Dutch book strategy takes as input the utility function and the agent's actual preferences, and gives as output the probability function. If this were true, it would make the MEU rule fairly

useless as a guide to action. But in fact it is not true. For we have seen that (according to the Dutch book argument) not any old set of preferences counts as rational: those which cause the agent to lose money are deemed irrational, and give rise to no (consistent) probability function. That is, the Dutch book argument is a constraint on rational preferences: it says that agents with certain preferences are rationally committed to having other preferences.

And this is the basic structure of decision theory. It is not in general concerned with the rationality or otherwise of any single decision by an agent: if someone wants to jump into an alligator pit, we might deem them irrational but decision theory will not. But someone who prefers jumping into alligator pits to lying on the beach, and prefers lying on the beach to jumping into snake pits, is constrained to prefer alligator pits to snake pits.

From this perspective, what decision theory aims to do is to state general, reasonable principles of rationality and use those principles to prove that any agent conforming to those principles must be behaving as if he is using MEU with respect to some probability measure and some utility function. The Dutch book argument can be understood as a rudimentary sort of decision theory, but so far it is at most vaguely formulated and its constraints on rationality are questionable (is it *per se* irrational to choose a course of action which always loses money?). We will see, in the rest of this chapter, how to do very much better.

B.2 Synchronic decision problems

To make more progress, we need to provide a formal mathematical framework for decision theory. We will eventually look at several such frameworks, but our first is fairly minimal. It consists of two parts: the possible outcomes of some action (usually called *events*), and the *rewards* which can accrue to an agent if he makes a certain bet on what the outcomes are.

For the moment, we can take the rewards to be elements of any set we like. The events have rather more structure, though: if A and B are events, so should be (A-and-B), (A-or-B), and (not-A). There is a natural mathematical way to represent something like this: recall that a *Boolean algebra* is a set equipped with a 'maximum' element 1, a 'minimum' element 0, associative and symmetric binary operations $\vee$ and $\wedge$, and a unary 'complement' operation $E \to \neg E$, such that:

1. $\vee$ and $\wedge$ are distributive over one another;
2. $E \vee (E \wedge F) = E \wedge (E \vee F) = E$;