# CHAPTER 16

# MAKING SIMPLE DECISIONS

*In which we see how an agent should make decisions so that it gets what it wants in an uncertain world—at least as much as possible and on average.*

In this chapter, we fill in the details of how utility theory combines with probability theory to yield a decision-theoretic agent—an agent that can make rational decisions based on what it believes and what it wants. Such an agent can make decisions in contexts in which uncertainty and conflicting goals leave a logical agent with no way to decide. A goal-based agent has a binary distinction between good (goal) and bad (non-goal) states, while a decision-theoretic agent assigns a continuous range of values to states, and thus can more easily choose a better state even when no best state is available.

Section 16.1 introduces the basic principle of decision theory: the maximization of expected utility. Section 16.2 shows that the behavior of a rational agent can be modeled by maximizing a utility function. Section 16.3 discusses the nature of utility functions in more detail, and in particular their relation to individual quantities such as money. Section 16.4 shows how to handle utility functions that depend on several quantities. In Section 16.5, we describe the implementation of decision-making systems. In particular, we introduce a formalism called a **decision network** (also known as an **influence diagram**) that extends Bayesian networks by incorporating actions and utilities. Section 16.6 shows how a decision-theoretic agent can calculate the value of acquiring new information to improve its decisions.

While Sections 16.1–16.6 assume that the agent operates with a given, known utility function, Section 16.7 relaxes this assumption. We discuss the consequences of preference uncertainty on the part of the machine—the most important of which is deference to humans.

## 16.1 Combining Beliefs and Desires under Uncertainty

We begin with an agent that, like all agents, has to make a decision. It has available some actions $a$. There may be uncertainty about the current state, so we'll assume that the agent assigns a probability $P(s)$ to each possible current state $s$. There may also be uncertainty about the action outcomes; the transition model is given by $P(s' \mid s, a)$, the probability that action $a$ in state $s$ reaches state $s'$. Because we're primarily interested in the outcome $s'$, we'll also use the abbreviated notation $P(\textsc{Result}(a) = s')$, the probability of reaching $s'$ by doing $a$ in the current state, whatever that is. The two are related as follows:

$$P(\textsc{Result}(a) = s') = \sum_s P(s) P(s' \mid s, a)\,.$$

Decision theory, in its simplest form, deals with choosing among actions based on the desirability of their *immediate* outcomes; that is, the environment is assumed to be episodic in the