Chapter 7

Decision Theory

Up to this point most of our discussion has been about epistemology. But probability theory originated in attempts to understand games of chance, and historically its most extensive application has been to practical decision-making. The Bayesian theory of probabilistic credence is a central element of decision theory, which developed throughout the twentieth century in philosophy, psychology, and economics. **Decision theory** searches for rational principles to evaluate the various acts available to an agent at any given moment. Given what she values (her utilities) and how she sees the world (her credences), decision theory recommends the act that is most efficacious for achieving those values from her point of view.

Decision theory has always been a crucial application of Bayesian theory. In his seminal *The Foundations of Statistics*, L.J. Savage wrote,

> Much as I hope that the notion of probability defined here is consistent with ordinary usage, it should be judged by the contribution it makes to the theory of decision. (Savage 1954, p. 27)

Decision theory has also been extensively studied, and a number of excellent book-length introductions are now available. (I recommend one in the Further Readings section of this chapter.) As a result, I have not attempted to make this chapter nearly so comprehensive as the preceding chapter on confirmation. I aim only to acquaint the reader with the main ideas and terminology one would need to work farther into the philosophy of decision theory, as well as the concepts we will need for discussions later in the book.

We will begin with the general mathematical notion of an expectation, followed by the philosophical notion of utility. We will then see how Savage

calculates expected utilities to determine rational preferences among acts, and the formal properties of rational preference that result. Next comes Richard Jeffery's Evidential Decision Theory, which improves on Savage by applying to probabilistically dependent states and acts. We will then discuss Jeffrey's troubles with certain kinds of risk-aversion (especially the Allais Paradox), and with Newcomb's Problem. Causal Decision Theory will be suggested as a better response to Newcomb. I will end by briefly tracing some of the historical back-and-forth about which decision theory handles Newcomb's problem best.

7.1 Calculating expectations

Suppose there's a numerical quantity—the number of hits a particular batter will have in tonight's game, say—and you have opinions about what value that quantity will take. We can then calculate your **expectation** for the quantity. While there are subtleties we will return to later, the basic idea of an expectation is to multiply each value the quantity might take by your credence that it'll take that value, then add up the results. So if you're 30% confident the batter will have 1 hit, 20% confident she'll have 2 hits, and 50% confident she'll have 3, your expectation for the number of hits is

$$0.30 \cdot 1 + 0.20 \cdot 2 + 0.50 \cdot 3 = 2.2 \tag{7.1}$$

Your expectation of a quantity is *not* the value you anticipate the quantity will actually take, or even the value you think it's most probable the quantity will take—in the baseball example, you're certain the batter won't have 2.2 hits in tonight's game! Your expectation of a quantity is more like an *estimate* of the value the quantity will take. When you're uncertain about the value of a quantity, a good estimate may straddle the line between multiple options.

We can also think of your expectation for a quantity as a weighted average of its possible values, with weights provided by your unconditional credences in those various possibilities. This weighted average becomes an actual average for repeatable situations. Suppose you're certain that this batter will play in many games over time. The **law of large numbers** says that if you satisfy the probability axioms, you'll have credence 1 that as the number of games increases, the average number of hits per game will tend towards your expectation for that quantity. In other words, you're highly confident that as the number of games approaches the limit, the batter's average hits per game will approach 2.2.[1] So while your expectation isn't

the number of hits you anticipate will actually happen in a given game, it is the *average* number of hits per game you expect in the long run.

We've already calculated expectations for a few different quantities in this book. For example, when you lack inadmissible evidence the Principal Principle requires your credence in a proposition to equal your expectation of its chance. (See especially our calculation in Equation (5.7).) But by far the most commonly calculated expectations in life are monetary values. For example, suppose you have the opportunity to buy stock in a company just before it announces quarterly earnings. If the announcement is good you'll be able to sell shares at \$100 each, but if the announcement is bad you'll be forced to sell at \$10 apiece. The value you place in these shares depends on your confidence in a good report. If you're 40% confident in a good earnings report, your expected value for each share is

$$\$100 \cdot 0.40 + \$10 \cdot 0.60 = \$46 \tag{7.2}$$

As a convention, we let positive monetary values stand for money accrued to the agent; negative monetary values are amounts the agent pays out. So your expectation of how much money you will receive for each share is \$46.

An agent's **fair price** for an investment is what she takes to be that investment's break-even point—she'd pay anything *up to* that amount of money in exchange for the investment. If you use expected values to make your investment decisions, your fair price for each share of the stock just described will be \$46. Given the opportunity to buy shares for less than \$46 each, you'll expect to make a profit on them; on the other hand, you'd expect to lose money on shares priced higher than that.

There are a couple of reasons why it's sensible to set your fair price for an investment equal to your expectation of its monetary return. First, suppose you know you're going to be confronted with this exact investment situation many, many times. The law of large numbers says that in the long run you should anticipate an average return of \$46 per share. So if you're going to adopt a standing policy for buying and selling such investments, you are highly confident that any price higher than \$46 will lose you money and any price lower than \$46 will make you money in the long-term. Second, expectations vary in intuitive ways when conditions change. If you become more confident in a good earnings report, you should be willing to pay more for a share, and the expected value reflects that. On the other hand, if you learn that a good earnings report will send the share value to only \$50, this decreases the expected value and also should decrease the price you'd be willing to pay.

An investment is a type of bet, and fair betting prices play a significant role in Bayesian lore. (We'll see one reason why in Chapter 9.) A bet that pays \$1 if proposition P is true and nothing otherwise has an expected value of

$$\$1 \cdot \mathrm{cr}(P) + \$0 \cdot \mathrm{cr}(\sim P) = \$\mathrm{cr}(P) \tag{7.3}$$

If you use expectations to calculate fair betting prices, your price for a gamble that pays \$1 on P equals your unconditional credence in P.

A lottery ticket is a common type of bet, and in the right situation calculating its expected value can be extremely lucrative. Ellenberg (2014, Ch. 11) relates the story of Massachusetts' Cash WinFall state lottery game, which was structured so that if the jackpot rose above a particular level, payoffs went up *even for players who won a prize other than the jackpot*. Because of this structure, when the jackpot rose high enough the expected payoff for a single ticket could climb higher than the price the state charged for that ticket. For example, on February 7, 2005 the expected value of a \$2 lottery ticket was \$5.53. The implications of this arrangement were understood by three groups of individuals—led respectively by an MIT student, a medical researcher in Boston, and a retiree in Michigan who had played a short-lived similar game in his home state. Of course, the expected value of a ticket isn't necessarily what you will win if you buy a single ticket, but because of the long-run behavior of expectations your confidence in a net profit goes up the more tickets you buy. So these groups bought a *lot* of tickets. For instance, on August 13, 2010 the MIT group bought around 700,000 tickets, almost 90% of the Cash WinFall tickets purchased that day. Their \$1.4 million investment netted about \$2.1 million in payouts, for a 50% profit in one day. Expected value theory can be *very* effective.

7.1.1 The move to utility

Yet sometimes we value things other than money. For example, suppose it's late at night, it's cold out, you're trying to catch a bus that costs exactly \$1, and you've got no money on you. A stranger offers either to give you \$1 straight up, or to flip a fair coin and give you \$2.02 if it comes up heads. It might be highly rational for you to prefer the guaranteed dollar even though its expected monetary value is less than that of the coin bet.

Decision theorists and economists explain this preference by introducing the notion of **utility**. Introduced by Daniel Bernoulli and Gabriel Cramer in the 18th century,[2] utility is a numerical quantity meant to directly measure how much an agent values an arrangement of the world. Just as we suppose that each agent has her own credence distribution, we will suppose that each

agent has a utility distribution over the propositions in language $\mathcal{L}$. The utility an agent assigns to a proposition represents how much she values that proposition's being true (or if you like, how happy that proposition's being true would make her). If an agent would be just as happy for one proposition to be true as another, she assigns them equal utility. But if it would make her happier for one of those propositions to be true, she assigns it the higher utility of the two.

While there continue to be debates between Subjective Bayesians and Objective Bayesians (in both the senses identified in Section 5.1.2) concerning probability and credence, almost everyone working on decision theory these days is a subjective utility theorist: utility distributions are assumed to be features of individual agents that may differ without the implication of irrationality. If I assign more value to the proposition that the Yankees win this year's pennant than the proposition that the Mets do, while you assign the opposite, neither of us need be irrational.

Utilities provide a uniform value-measurement scale. To see what I mean, consider the fact that in the bus example above, you don't value each dollar equally. Going from zero dollars to one dollar would mean a lot to you; it would get you out of the cold and on your way home. Going from one dollar to two dollars would not mean nearly as much in your present context. Not every dollar represents the same amount of value in your hands, so counting the number of dollars in your possession is not a consistent measure of how much you value your current state. On the other hand, utilities measure value uniformly. We stipulate that each added unit of utility (sometimes called a **util**) is equally valuable to an agent. She is just as happy to go from -50 utils to -49 as she is to go from 1 util to 2, and so on.

Having introduced this uniform value scale, we can explain your preferences in the bus case using expectations. Admittedly, the coin flip gamble has a higher expected *monetary* payoff ($1.01) than the guaranteed dollar. But monetary value doesn't always translate neatly to utility, and utility relfects the values on which you truly make your decisions. Let's say receiving one dollar is worth 100 utils to you in this case, while receiving $2.02 is worth 102 utils. (The larger amount of money is still more valuable to you; it just isn't more valuable by much.) Now when we calculate the expected *utility* of the gamble, it only comes to 51 utils, which is much less than the 100 expected utils associated with the guaranteed dollar. So you prefer the dollar guarantee.

The setup of this example is somewhat artificial, because it makes the value of money change radically at a particular cutoff point. But economists think money generally has a **decreasing marginal utility** for agents.

While an agent always receives some positive utility from receiving another dollar (or peso, or yuan, or...), the more dollars she already has the less that additional bit of utility will be. The first billion you earn makes your family comfortable; the second billion doesn't have as much significance in your life. Postulating an underlying locus of value distinguishable from net worth helps explain why we don't always chase the next dollar as hard as we chased the first.

With that said, quantifying value on a constant numerical scale introduces many of the same problems we found with quantifying confidence. First, it's not clear that a real agent's psychology will always be as rich as a numerical utility structure seems to imply. And second, the moment you assign numerical utilities to every arrangement of the world you make them all comparable; the possibility of incommensurable values is lost. (Compare Section 1.2.2.)

7.2 Expected Utility Theory

7.2.1 Preference orderings, and money pumps

A **decision problem** presents an agent with a partition of **acts**, from which she must choose exactly one. If the agent is certain how much utility will be generated by the performance of each act, the choice is simple—she prefers the act leading to the highest-utility result. Yet the utility resulting from an act often depends on features of the world beyond the agent's control (think, for instance, of the factors determining whether a particular career choice turns out well), and the agent may be uncertain how those features stand. In that case, the agent needs a technique for factoring her uncertainties into her decision.

There are many lively controversies in decision theory, but we will focus mainly on the question of how an agent should combine her credences and utilities to determine her preferences among acts. The first assumption of decision theory is that a rational agent uses some **valuation function** to assign each act a numerical score. This creates an ordering over the acts; the agent prefers act A to act B just in case A receives a higher score. In that case we write $A \succ B$. If A and B receive exactly the same score, the agent is indifferent between A and B and we write $A \sim B$. Given a particular decision problem, a rational agent will select the available act that she most prefers (or—if there are ties at the top—an act from amongst those she most prefers).

The moment we require an agent to set her preferences according to

a numerical score, we introduce a number of formal properties into her preference ordering. For example:

Preference Asymmetry: There do not exist acts A and B such that the agent both prefers A to B and prefers B to A.

Preference Transitivity: For any acts A, B, and C, if the agent prefers A to B and B to C, then the agent prefers A to C.

Why must these properties hold? Take Preference Asymmetry. The agent assigns $A \succ B$ (the agent prefers A to B) just in case her valuation function gives A a higher score. In that case B will receive the lower score, so it won't be the case that $B \succ A$. (The reader may construct a similar argument for Preference Transitivity.)

Hopefully it seems sensible that a rational preference ordering should satisfy these properties. One might object to Preference Transitivity that an agent may prefer A to B and prefer B to C, but never have thought to compare A to C. In other words, one might think that an agent's preference ordering could go silent on the comparison between A and C. Yet once more, having a numerical valuation function over the entire set of acts settles this issue; it forces the agent's preferences to form a total ordering. Decision theorists sometimes express this as:

Preference Completeness: For any acts A and B, exactly one of the following is true: the agent prefers A to B, the agent prefers B to A, or the agent is indifferent between the two.

Given that we're dealing with complete preferences, we can go beyond our intuitions about Preference Asymmetry and Preference Transitivity to provide an *argument* for the two. Consider a situation in which some of us find ourselves all the time. On any given weeknight, I would prefer to do something else over washing the dishes. (Going to a movie? Great! Watching the game? Good idea!) But when the week ends and the dishes have piled up, I realize that I would've preferred foregoing one of those weeknight activites in order to avoid a disgusting kitchen. Each of my individual decisions was made in accordance with my preferences among the acts I was choosing between at the time, yet together those local preferences add up to a global outcome I disprefer.

A student once suggested to me that he prefers eating out to cooking for himself, prefers eating at a friend's to eating out, but prefers his own cooking to his friend's. Imagine one night my student is preparing himself dinner, then decides he'd prefer to order out. He calls up the takeout place,

but before they pick up the phone he decides he'd rather drive to his friend's for dinner. He gets in his car and is halfway to his friend's, when he decides he'd rather cook for himself. At which point he turns around and goes home, having wasted a great deal of time and energy. Each of those choices reflects the student's preference between the two options he considers at the time, yet their net effect is to leave him right back where he started meal-wise and out a great deal of effort overall.

My student's preferences violate Transitivity; as a result he's susceptible to a **money pump**. In general, a money pump against intransitive preferences (preferring A to B, B to C, and C to A) can be constructed like this: Suppose you're about to perform act B, and I suggest I could make it possible to do A instead. Since you prefer A to B, there must be *some* amount of something (we'll just suppose it's money) you'd be willing to pay me for the option to perform A. So you pay the price, are about to perform A, but then I hold out the possibility of performing C instead. Since you prefer C to A, you pay me a small amount to make that switch. But then I offer you the opportunity to perform B rather than C—for a small price, of course. And now you're back to where you started with respect to A, B, and C, but out a few dollars for your trouble. To add insult to injury, I could repeat this set of trades again, and again, milking more and more money out of you until I decide to stop. Hence the "money pump" terminology.

Violating Preference Transitivity leaves one susceptible to such sets of money-pumping trades. (If you violate Preference Asymmetry, the money pump is even simpler.) In a money pump, the agent proceeds through a number of exchanges, each of which looks favorable given his preferences between the two actions involved. But when those exchanges are combined, the total package produces a net loss (which the agent would prefer to avoid). The money pump therefore seems to reveal an internal inconsistency between the agent's local and global preferences, as in my dishwashing example. (We will further explore this kind of inconsistency in our Chapter 9 discussion of Dutch Books.) The irrationality of being susceptible to a money pump has been taken as a strong argument against violating Preference Asymmetry or Transitivity.

7.2.2 Savage's expected utility

Savage (1954) frames decision problems using a partition of acts available to the agent and a partition of **states** the world might be in. A particular act performed with the world in a particular state produces a particular **outcome**. Agents assign numerical utility values to outcomes; given partial

information they also assign credences over states.[3]

Here's a simple example: Suppose you're trying to decide whether to carry an umbrella today. This table displays the utilities you assign various outcomes:

	rain	dry
take umbrella	0	−1
leave it	−10	0

You have two available acts, represented in the rows of the table. There are two possible states of the world, represented in the columns. Performing a particular act when the world is in a particular state produces a particular outcome. If you leave your umbrella behind and it rains, the outcome is you walking around wet. The cells in the table report your utilities for the various possible outcomes. Your utility for walking around wet is −10 utils, while carrying an umbrella on a dry day is inconvenient but not nearly as unpleasant (−1 util).

Now suppose you're uncertain about the state of the world; you have a 0.30 credence in rain. How can you evaluate the two available acts and set your preferences between them? For a finite partition $S_1, S_2, \ldots, S_n$ of possible states of the world, Savage endorses the following valuation function:

$$\begin{aligned} \mathrm{EU}_{\mathrm{SAV}}(A) = \mathrm{u}(A \,\&\, S_1) \cdot \mathrm{cr}(S_1) + \mathrm{u}(A \,\&\, S_2) \cdot \mathrm{cr}(S_2) \\ + \ldots + \mathrm{u}(A \,\&\, S_n) \cdot \mathrm{cr}(S_n) \end{aligned} \tag{7.4}$$

Here A is the particular act being evaluated. Savage evaluates acts by calculating their expected utilities; $\mathrm{EU}_{\mathrm{SAV}}(A)$ represents the expected utility of act A calculated in the manner Savage prefers. (We'll see other ways of calculating expected utility later on.) $\mathrm{cr}(S_i)$ is the agent's unconditional credence that the world is in state S_i; $\mathrm{u}(A \,\&\, S_i)$ is the utility she assigns to the outcome that will eventuate should she perform act A in state S_i. So $\mathrm{EU}_{\mathrm{SAV}}$ calculates the weighted average of the utilities the agent might receive if she performs A, weighted by her credence that she will receive each one. Savage holds that given a partition of acts to consider, a rational individual will prefer to perform an act with at least as great an expected utility as that of any act on offer.

What does that mean for the present case? We calculate expected utilities for each of the acts available as follows:

$$\begin{aligned} \mathrm{EU}_{\mathrm{SAV}}(\text{take}) &= 0 \cdot 0.30 + -1 \cdot 0.70 = -0.7 \\ \mathrm{EU}_{\mathrm{SAV}}(\text{leave}) &= -10 \cdot 0.30 + 0 \cdot 0.70 = -3 \end{aligned} \tag{7.5}$$

Taking the umbrella has the higher expected utility, so Savage thinks that if you're rational you'll prefer to take the umbrella. You're more confident it'll be dry than rain, but this is outweighed by the much greater disutility of a disadvantageous decision in the latter case than the former.

$\mathrm{EU}_{\mathrm{SAV}}$ is a valuation function that combines credences and utilities in a specific way to assign numerical scores to acts. As a numerical valuation function, it generates a preference ordering satisfying Asymmetry, Transitivity, and Completeness. But calculating expected utilities this way also introduces new features not shared by all valuation functions. For example, Savage's expected utility theory yields preferences that satisfy the:

Dominance Principle: If act A produces a higher-utility outcome than act B in each possible state of the world, then A is preferred to B.

The Dominance Principle[4] seems intuitively like a good rational principle. Yet (surprisingly) there are decision problems in which it gives very bad advice. Since Savage's expected utility theory entails the Dominance Principle, it can be relied upon only when we don't find ourselves in decision problems like that.

7.2.3 Jeffrey's theory

To see what can go wrong with dominance reasoning, consider this example from (Weirich 2012):

> A student is considering whether to study for an exam. He reasons that if he will pass the exam, then studying is wasted effort. Also, if he will not pass the exam, then studying is wasted effort. He concludes that because whatever will happen, studying is wasted effort, it is better not to study.

The student entertains two possible acts—study or not study—and two possible states of the world—he either passes the exam or he doesn't. His utility table looks something like this:

	pass	fail
study	18	−5
don't study	20	−3

Because studying costs effort, passing having not studied is better than passing having studied, and failing having not studied is also better than failing having studied. So whether he passes or fails, not studying yields a higher utility. By the Dominance Principle, the student should prefer not studying to studying.

This is clearly a horrible argument; it ignores the fact that whether the student studies *affects whether he passes the exam*.[5] The Dominance Principle—and Savage's expected utility theory in general—breaks down when the state of the world that eventuates depends on the act the agent performs. Savage recognizes this limitation, and so requires that the acts and states used in framing decision problems be independent of each other. Jeffrey (1965), however, notes that in real life we often analyze decision problems in terms of dependent acts and states. Moreover, he worries that agents might face decision problems in which they are unable to identify independent acts and states.[6] So it would be helpful to have a decision theory that didn't require acts and states to be independent.

Jeffrey offers just such a theory. The key innovation is a new valuation function that calculates expected utilities differently from Savage's. Given an act A and a finite partition $S_1, S_2, \ldots, S_n$ of possible states of the world,[7] Jeffrey calculates

$$\begin{aligned} \mathrm{EU}_{\mathrm{EDT}}(A) = \mathrm{u}(A \,\&\, S_1) \cdot \mathrm{cr}(S_1 \,|\, A) + \mathrm{u}(A \,\&\, S_2) \cdot \mathrm{cr}(S_2 \,|\, A) \\ + \ldots + \mathrm{u}(A \,\&\, S_n) \cdot \mathrm{cr}(S_n \,|\, A) \end{aligned} \tag{7.6}$$

I'll explain the "EDT" subscript later on; for now, it's crucial to see that Jeffrey alters Savage's approach (Equation (7.4)) by replacing the agent's *unconditional* credence that a given state S_i obtains with the agent's *conditional* credence that S_i obtains given A. This incorporates the possibility that performing the act the agent is evaluating will change the probabilities of various states of the world.

To see how this works, consider Jeffrey's (typically civilized) example of a guest deciding whether to bring white or red wine to dinner. The guest is certain his host will serve either chicken or beef, but doesn't know which. The guest's utility table is as follows:

	chicken	beef
white	1	−1
red	0	1

For this guest, bringing the right wine is always pleasurable. Red wine with chicken is merely awkward, while white wine with beef is a disaster.

Typically, the entree for an evening is settled well before the guests arrive. But let's suppose our guest suspects his host is especially accommodating. The guest is 75% confident that the host will select a meat in response to the wine provided. (Perhaps the host has a stocked pantry, and waits to prepare dinner until the wine has arrived.) In that case, the state (meat served) depends on the agent's act (wine chosen). This means the agent cannot assign a uniform unconditional credence to each state prior to his decision. Instead, the guest assigns one credence to chicken conditional on his bringing white, and another credence to chicken conditional on his bringing red. These credences are reflected in the following table:

	chicken	beef
white	0.75	0.25
red	0.25	0.75

It's important to read the credence table differently from the utility table. In the utility table, the entry in the white/chicken cell is the agent's utility assigned to the outcome of chicken served *and* white wine. In the credence table, the white/chicken entry is the agent's credence in chicken served *given* white wine. The probability axioms and Ratio Formula demand that all the credences conditional on white wine sum to 1, so the values in the first row sum to 1 (as do the values in the second row).

We can now use Jeffrey's formula to calculate the agent's expected utility for each act. For instance,

$$\begin{aligned} \mathrm{EU}_{\mathrm{EDT}}(\text{white}) &= \mathrm{u}(\text{white} \,\&\, \text{chicken}) \cdot \mathrm{cr}(\text{chicken} \,|\, \text{white}) \\ &\qquad + \mathrm{u}(\text{white} \,\&\, \text{beef}) \cdot \mathrm{cr}(\text{beef} \,|\, \text{white}) \\ &= 1 \cdot 0.75 + -1 \cdot 0.25 \\ &= 0.5 \end{aligned} \tag{7.7}$$

(We multiply the values in the first row of the utility table by the corresponding values in the first row of the credence table, then sum the results.) A similar calculation yields $\mathrm{EU}_{\mathrm{EDT}}(\text{red}) = 0.75$. Bringing red wine has a higher expected utility for the agent than bringing white, so the agent should prefer bringing red.

Earlier I said somewhat vaguely that Savage requires acts and states to be "independent"; Jeffrey's theory gives that notion a precise meaning. $\mathrm{EU}_{\mathrm{EDT}}$ revolves around an agent's conditional credences, so for Jeffrey the relevant notion of independence is probabilistic independence relative to the agent's credence function. That is, an act A and state S_i are independent

for Jeffrey just in case

$$\mathrm{cr}(S_i \mid A) = \mathrm{cr}(S_i) \tag{7.8}$$

In the special case where the act A being evaluated is independent of each state S_i, the $\mathrm{cr}(S_i \mid A)$ expressions in Jeffrey's formula may be replaced with $\mathrm{cr}(S_i)$ expressions. This makes Jeffrey's expected utility calculation identical to Savage's. When acts and states are probabilistically independent, Jeffrey's theory yields the same preferences as Savage's. And since Savage's theory entails the Dominance Principle, Jeffrey's theory will also embrace Dominance in this special case.

But what happens to Dominance when acts and states are *dependent*? Here Jeffrey offers a nuclear deterrence example. Suppose a nation is choosing whether to arm itself with nuclear weapons, and knows its rival nation will follow its lead. The possible states of the world under consideration are war versus peace. The utility table might be:

	war	peace
arm	−100	0
disarm	−50	50

Wars are worse when both sides have nuclear arms; peace is also better without nukes on hand (because of nuclear accidents, etc.). A dominance argument is now available since whichever state obtains, disarming provides the greater utility. So applying Savage's theory to this example would yield a preference for disarming.

Yet an advocate of nuclear deterrence will take the states in this example to depend on the acts. The deterrence advocate's credence table might be:

	war	peace
arm	0.1	0.9
disarm	0.8	0.2

The idea of deterrence is that if both countries have nuclear arms, war becomes much less likely. If arming increases the probability of peace, the acts and states in this example are probabilistically dependent. Jeffrey's theory calculates the following expected utilities from these tables:

$$\begin{aligned} \mathrm{EU}_{\mathrm{EDT}}(\text{arm}) &= -100 \cdot 0.1 + 0 \cdot 0.9 = -10 \\ \mathrm{EU}_{\mathrm{EDT}}(\text{disarm}) &= -50 \cdot 0.8 + 50 \cdot 0.2 = -30 \end{aligned} \tag{7.9}$$

Relative to the deterrence advocate's credences, Jeffrey's theory yields a preference for arming. Act/state dependence has created a preference ordering at odds with the Dominance Principle.[8] When an agent takes the acts

and states in a decision problem to be independent, Jeffrey's and Savage's decision theories are interchangeable, and dominance reasoning can be relied upon. But Jeffrey's theory also provides reliable verdicts when acts and states are dependent, a case in which Savage's theory and the Dominance Principle may fail.

7.2.4 Risk aversion, and Allais' paradox

People sometimes behave strangely when it comes to taking risks. Many agents are **risk-averse**; they would rather have a sure \$10 than take a 50-50 gamble on \$30, even though the expected dollar value of the latter is greater than that of the former.

Economists have traditionally explained this preference by appealing to the declining marginal utility of money. If the first \$10 yields much more utility than the next \$20 for the agent, then the sure \$10 may in fact have a higher expected *utility* than the 50-50 gamble. This makes the apparently risk-averse behavior perfectly rational. But it does so by portraying the agent as only *apparently* risk-averse. The suggestion is that the agent would be happy to take a risk if only it offered him a higher expectation of what he really values—utility. But might some agents be genuinely willing to give up a bit of expected utility if it meant they didn't have to gamble? If we could offer agents a direct choice between a guaranteed 10 utils and a 50-50 gamble on 30, might some prefer the former? (Recall that utils are defined so as not to decrease in marginal value.) And might that preference be rationally permissible?

Let's grant for the sake of argument that risk-aversion concerning monetary gambles can be explained by attributing to the agent a decreasing marginal utility distribution over dollars. Other documented responses to risk cannot be explained by *any* kind of utility distribution. Suppose a fair lottery is to be held with 100 numbered tickets. You are offered two gambles to choose between, with the following payoffs should particular tickets be drawn:

	Ticket 1	Tickets 2–11	Tickets 12–100
Gamble A	\$1M	\$1M	\$1M
Gamble B	\$0	\$5M	\$1M

Which gamble would you prefer? After recording your answer somewhere, consider the next two gambles (on the same lottery) and decide which of them you would prefer if they were your only options:

	Ticket 1	Tickets 2–11	Tickets 12–100
Gamble C	\$1M	\$1M	\$0
Gamble D	\$0	\$5M	\$0

When subjects are surveyed, they often prefer Gamble D to C; they're probably not going to win anything, but if they do they'd like a serious shot at \$5 million. On the other hand, many of the same subjects prefer Gamble A to B, because A guarantees them a payout of \$1 million.

Yet anyone who prefers A to B while at the same time preferring D to C violates Savage's[9]

Sure-Thing Principle: If two acts yield the same outcome on a particular state, any preference between them remains the same if that outcome is changed.

In our example, Gambles A and B yield the same outcome for tickets 12 through 100: 1 million dollars. If we change that common outcome to 0 dollars, we get Gambles C and D. The Sure-Thing Principle requires an agent who prefers A to B also to prefer C to D. Put another way: if the Sure-Thing Principle holds, we can determine a rational agent's preferences between any two acts by focusing exclusively on the states for which those acts produce different outcomes. In both the decision problems here, tickets 12 through 100 produce the same outcome no matter which act the agent selects. So we ought to be able to determine her preferences by focusing exclusively on the outcomes for tickets 1 through 11. Yet if we focus exclusively on those tickets, A stands to B in exactly the same relationship as C stands to D. So the agent's preferences across the two decisions should be aligned.

The Sure-Thing Principle is a theorem of Savage's decision theory. It is also therefore a theorem of Jeffrey's decision theory for cases in which acts and states are independent, as they are in the present gambling example. Thus preferring A to B while preferring D to C—as real-life subjects often do—is incompatible with these two decision theories. And here we can't chalk up the problem to working with dollars rather than utils. There is no possible utility distribution over dollars on which Gamble A has a higher expected utility than Gamble B while Gamble D has a higher expected utility than Gamble C. (See Exercise 7.5.)

Jeffrey and Savage, then, must shrug off these commonly-paired preferences as irrational. Yet Maurice Allais, the Nobel-winning economist who introduced the gambles in his (1953), thought that both sets of preferences could be perfectly rational, and rationally held together. Because it's impossible to maintain these seemingly-reasonable preferences while hewing to

standard decision theory, the example is now known as Allais' Paradox. Allais thought the example revealed a deep flaw in the decision theories we've been considering.

We have been discussing these decision theories as *normative* accounts of how *rational* agents behave. Economists, however, often assume that decision theory provides an accurate *descriptive* account of *real* agents' market decisions. Real-life subjects' responses to cases like the Allais Paradox prompted economists to develop new descriptive theories of agents' behavior, such as Kahneman and Tversky's Prospect Theory (Kahneman and Tversky 1979; Tversky and Kahneman 1992). More recently, Buchak (2013) has proposed a generalization of standard decision theory that accounts for risk aversion without positing declining marginal utilities and is consistent with the Allais preferences subjects often display.

7.3 Causal Decision Theory

Although we have been focusing on the expected values of propositions describing acts, Jeffrey's valuation function can be applied to any sort of proposition. For example, suppose my favorite player has been out of commission for weeks with an injury, and I am waiting to hear whether he will play in tonight's game. I start wondering whether I would prefer that he play tonight on not. Usually it would make me happy to see him on the field, but there's the possibility that he will play despite his injury's not being fully healed. That would definitely be a bad outcome. So now I combine my credences about states of the world (is he fully healed? is he not?) with my utilities for the various possible outcomes (plays fully healed, plays not fully healed, etc.) to determine how happy I would be to hear that he's playing or not playing. Having calculated expected utilities for both "plays" and "doesn't play", I decide whether I'd prefer that he play or not.

Put another way, I can use Jeffrey's expected utility theory to determine whether I would consider it good news or bad were I to hear that my favorite player will be playing tonight. And I can do so whether or not I have *any* influence on the truth of that proposition. Jeffrey's theory is sometimes described as calculating the "news value" of a proposition.

Even for propositions describing our own acts, Jeffrey's expected utility calculation assesses news value. I might be given a choice between a sure \$1 and a 50-50 chance of \$2.02. I would use my credences and utility function to determine expected values for each act, then declare which option I preferred. But notice that this calculation would go exactly the same if instead

of my selecting among the options, someone else was selecting on my behalf. If my utility function assigns declining marginal utility to money, I might prefer just as much that someone else pick the sure dollar for me as I would prefer picking that option for myself. What's ultimately being compared are the proposition *that I receive a sure dollar* and the proposition *that I receive whatever payoff results from a particular gamble*. Whether I have the ability to make one of those propositions true rather than the other is irrelevant to Jeffrey's preference calculations.

7.3.1 Newcomb's Problem

Jeffrey's attention to news value irrespective of agency leads him into trouble with Newcomb's Problem. This problem was introduced to philosophy by Robert Nozick, who attributed its construction to the physicist William Newcomb. Here's how Nozick introduced the problem:

> Suppose a being in whose power to predict your choices you have enormous confidence. (One might tell a science-fiction story about a being from another planet, with an advanced technology and science, who you know to be friendly, etc.) You know that this being has often correctly predicted your choices in the past (and has never, so far as you know, made an incorrect prediction about your choices), and furthermore you know that this being has often correctly predicted the choices of other people, many of whom are similar to you, in the particular situation to be described below. One might tell a longer story, but all this leads you to believe that almost certainly this being's prediction about your choice in the situation to be discussed will be correct.
>
> There are two boxes. [The first box] contains $1,000. [The second box] contains either $1,000,000, or nothing. . . . You have a choice between two actions: (1) taking what is in both boxes (2) taking only what is in the second box.
>
> Furthermore, and you know this, the being knows that you know this, and so on:
>
> (I) If the being predicts you will take what is in both boxes, he does not put the $1,000,000 in the second box.
>
> (II) If the being predicts you will take only what is in the second box, he does put the $1,000,000 in the second box.

> The situation is as follows. First the being makes its prediction. Then it puts the $1,000,000 in the second box, or does not, depending upon what it has predicted. Then you make your choice. What do you do? (1969, pp. 114–5)

Historically, Newcomb's Problem prompted the development of a new kind of decision theory, now known as Causal Decision Theory (sometimes just "CDT"). At the time of Nozick's discussion, extant decision theories (such as Jeffrey's) seemed to recommend taking just one box in Newcomb's Problem (so-called "one-boxing"). But many philosophers thought two-boxing was the rational action.[10] Here's why: By the time you make your decision, the being has already made its prediction and taken its action. So the money is already either in the second box, or it's not—nothing you decide can affect whether the money is there. However much money is in the second box, you're going to get more money ($1,000 more) if you take both boxes. So you should two-box.

I've quoted Nozick's original presentation of the problem because in the great literature that has since grown up around Newcomb, there is often debate about what exactly counts as "a Newcomb Problem". Does it matter if the predictor is *perfect* at making predictions, or if the agent is *certain* that the prediction will be correct? Does it matter *how* the predictor makes its predictions, and whether backward causation (some sort of information fed backwards from the future) is involved? Perhaps more importantly, who *cares* about such a strange and fanciful problem?

But our purpose is not generalized Newcombology—we want to understand why Newcomb's Problem spurred the development of Causal Decision Theory. That can be understood by working with just one version of the problem. Or better yet, it can be understood by working with a kind of problem that comes up in everyday life, and is much less fanciful:

> I'm standing at the bar, trying to decide whether to order a third appletini. Drinking a third appletini is the kind of act much more typical of people with addictive personalities. People with addictive personalities also tend to become smokers. I'd kind of like to have another drink, but I *really* don't want to become a smoker (smoking causes lung-cancer, is increasingly frowned-upon in my social circle, etc.). So I shouldn't order that next appletini.

Let's work through the reasoning here on decision-theoretic grounds. First, stipulate that I have the following utility table:

	smoker	non
third appletini	−99	1
no more	−100	0

Ordering the third appletini is a dominant act. But dominance should dictate preference only when acts and states are independent, and my concern here is that they're not. My credence distribution has the following features (with A, S, and P representing the propositions that I order the appletini, that I become a smoker, and that I have an addictive personality, respectively):

$$\mathrm{cr}(S \mid P) > \mathrm{cr}(S \mid \sim P) \tag{7.10}$$
$$\mathrm{cr}(P \mid A) > \mathrm{cr}(P \mid \sim A) \tag{7.11}$$

I'm more confident I'll become a smoker if I have an addictive personality than if I don't. And having that third appletini is a positive indication that I have an addictive personality. Combining these two equations (and making a couple more assumptions I won't bother spelling out), we get:

$$\mathrm{cr}(S \mid A) > \mathrm{cr}(S \mid \sim A) \tag{7.12}$$

From my point of view, ordering the third appletini is positively correlated with becoming a smoker. Looking back at the utility table, I do not consider the states listed along the top to be probabilistically independent of the acts along the side. Now I calculate my Jeffrey expected utilities for the two acts:

$$\begin{aligned} \mathrm{EU}_{\mathrm{EDT}}(A) &= -99 \cdot \mathrm{cr}(S \mid A) + 1 \cdot \mathrm{cr}(\sim S \mid A) \\ \mathrm{EU}_{\mathrm{EDT}}(\sim A) &= -100 \cdot \mathrm{cr}(S \mid \sim A) + 0 \cdot \mathrm{cr}(\sim S \mid \sim A) \end{aligned} \tag{7.13}$$

Looking at these equations, you might think that A receives the higher expected utility. But I assign a considerably higher value to $\mathrm{cr}(S \mid A)$ than $\mathrm{cr}(S \mid \sim A)$, so the -99 in the top equation is multiplied by a significantly larger quantity than the -100 in the bottom equation. Assuming the correlation between S and A is strong enough, $\sim A$ receives the better expected utility and I prefer to perform $\sim A$.

But this is all wrong! Whether I have an addictive personality is (let's say) determined by genetic factors, not anything I could possibly affect at this point in my life. The die is cast (so to speak); I either have an addictive personality or I don't; it's already determined (in some sense) whether an addictive personality is going to lead me to become a smoker. Nothing

about this appletini—whether I order it or not—is going to change that. So I might as well enjoy the drink.

Assuming the reasoning in the previous paragraph is correct, it's an interesting question why Jeffrey's decision theory yields the wrong result. The answer is that on Jeffrey's theory ordering the appletini gets graded down because it would be bad news about my future. If I order the drink, that's evidence that I have an addictive personality (as indicated in Equation (7.11)), which is unfortunate because of its potential consequences for becoming a smoker. I expect a world in which I order that drink to be a worse world than a world in which I don't, and this is reflected in the $\mathrm{EU}_{\mathrm{EDT}}$ calculation. Jeffrey's theory assesses the act of ordering a third appletini not in terms of the consequences it will *cause* to come about, but instead in terms of the consequences it provides *evidence* will come about. For this reason Jeffrey's theory is described as an Evidential Decision Theory (or "EDT").

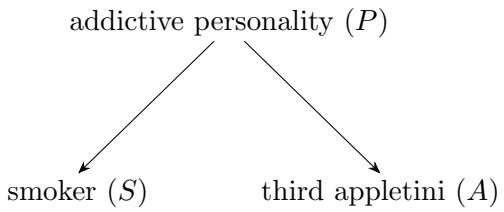
The trouble with Evidential Decision Theory is that an agent's performing an act may be *evidence* of a consequence that it's too late for her to *cause* (or *prevent*). Even though the act indicates the consequence, it seems irrational to factor the value of that consequence into a decision about whether to peform the act. As Skyrms (1980a, p. 129) puts it, my not having the third drink in order to avoiding becoming a smoker would be "a futile attempt to manipulate the cause by suppressing its symptoms." In making decisions we should attend to what we can control—to the causal consequences of our acts. Weirich writes,

> Deliberations should attend to an act's causal influence on a state rather than an act's evidence for a state. A good decision aims to produce a good outcome rather than evidence of a good outcome. It aims for the good and not just signs of the good. Often efficacy and auspiciousness go hand in hand. When they come apart, an agent should perform an efficacious act rather than an auspicious act. (2012)

7.3.2 A causal approach

The causal structure of our third drink example is depicted in Figure 7.1. As we saw in Chapter 3, correlation often indicates causation—but not *always*. Propositions on the tines of a causal fork will be probabilistically correlated even though neither causes the other. This accounts for A's being relevant to S on my credence function (Equation (7.12)) even though my ordering the third appletini has no causal influence on whether I'll become a smoker.

Figure 7.1: Third drink causal fork



The causally spurious correlation in my credences affects Jeffrey's expected utility calculation because that calculation works with credences in states conditional on acts ($\mathrm{cr}(S_i \mid A)$). Jeffrey replaced Savage's $\mathrm{cr}(S_i)$ with this conditional expression to track dependencies between states and acts. The Causal Decision Theorist responds that while credal correlation is a kind of probabilistic dependence, it may fail to track the causal dependences on which preferences should be based. So the Causal Decision Theorist's valuation function is:

$$\begin{aligned}\mathrm{EU}_{\mathrm{CDT}}(A) = \mathrm{u}(A \,\&\, S_1)\cdot\mathrm{cr}(A \mathrel{\square\!\!\rightarrow} S_1) + \mathrm{u}(A \,\&\, S_2)\cdot \mathrm{cr}(A \mathrel{\square\!\!\rightarrow} S_2) \\ + \ldots + \mathrm{u}(A \,\&\, S_n)\cdot \mathrm{cr}(A \mathrel{\square\!\!\rightarrow} S_n)\end{aligned} \tag{7.14}$$

Here $A \mathrel{\square\!\!\rightarrow} S$ represents the subjunctive conditional "If the agent were to perform act A, state S would occur."[11] Causal Decision Theory uses such conditionals to track causal relations in the world.[12] Of course, an agent may be uncertain what consequences a given act A would cause. So $\mathrm{EU}_{\mathrm{CDT}}$ looks across the partition of states $S_1, \ldots, S_n$ and invokes the agent's credence that A would cause any particular given S_i.

For many decision problems, Causal Decision Theory yields the same results as Evidential Decision Theory. In Jeffrey's wine example, it's plausible that

$$\mathrm{cr}(\text{chicken} \mid \text{white}) = \mathrm{cr}(\text{white} \mathrel{\square\!\!\rightarrow} \text{chicken}) = 0.75 \tag{7.15}$$

The guest's credence that chicken is served on the condition that she brings white wine is equal to her credence that if she were to bring white, chicken would be served. So one may be substituted for the other in expected utility calculations, and CDT's evaluations turn out the same as Jeffrey's.

But when conditional credences fail to track causal relations (as in cases with causal forks), the two theories may yield different results. This is in

part due to their differing notions of independence. EDT treats act A and state S as independent when they are *probabilistically* independent relative to the agent's credence function. CDT focuses on whether the agent takes A and S to be *causally* independent, which occurs just when

$$\mathrm{cr}(A \mathbin{\square\!\!\rightarrow} S) = \mathrm{cr}(S) \tag{7.16}$$

When A has no causal influence on S, the agent's credence that S will occur if she performs A is just her credence that S will occur. In the third drink example my ordering another appletini may be evidence that I'll become a smoker, but it has no causal bearing on whether I take up smoking. So from a Causal Decision Theory point of view, the acts and states in that problem are independent. When acts and states are independent, dominance reasoning is appropriate, so I should prefer the dominant act and order that third appletini.

Now we can return to a version of the Newcomb Problem that distinguishes Causal from Evidential Decision Theory. Suppose that the "being" in Nozick's story makes its prediction by analyzing your brain state prior to your making the decision and applying a complex neuro-psychological theory. The being's track record makes you 99% confident that its predictions will be correct. And to simplify matters, let's suppose you assign exactly 1 util to each dollar, no matter how many dollars you already have. Then your utility and credence matrices for the problem are:

Utilities

	P_1	P_2
T_1	1,000,000	0
T_2	1,001,000	1,000

Credences

	P_1	P_2
T_1	0.99	0.01
T_2	0.01	0.99

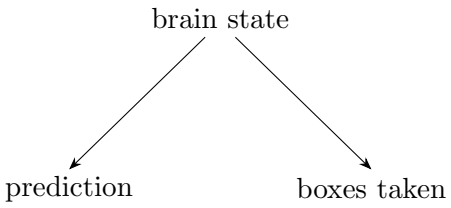
where T_1 and T_2 represent the acts of taking one box or two boxes (respectively), and P_1 and P_2 represent the states of what the being predicted.

Jeffrey calculates expected values for the acts as follows:

$$\begin{aligned}\mathrm{EU}_{\mathrm{EDT}}(T_1) &= \mathrm{u}(T_1 \,\&\, P_1) \cdot \mathrm{cr}(P_1 \,|\, T_1) + \mathrm{u}(T_1 \,\&\, P_2) \cdot \mathrm{cr}(P_2 \,|\, T_1) = 990,000 \\ \mathrm{EU}_{\mathrm{EDT}}(T_2) &= \mathrm{u}(T_2 \,\&\, P_1) \cdot \mathrm{cr}(P_1 \,|\, T_2) + \mathrm{u}(T_2 \,\&\, P_2) \cdot \mathrm{cr}(P_2 \,|\, T_2) = 11,000\end{aligned} \tag{7.17}$$

So Evidential Decision Theory recommends one-boxing. Yet we can see from Figure 7.2 that this version of the Newcomb Problem contains a causal fork;

Figure 7.2: Newcomb Problem causal fork



the being's prediction is based on your brain state, which also has a causal influence on the number of boxes you take. This should make us suspicious of EDT's recommendations. The agent's act and the being's prediction are probabilistically correlated in the agent's credences, as the credence table reveals. But that's not because the number of boxes taken has any causal influence on the prediction.

Causal Decision Theory calculates expected utilities in the example like this:

$$\begin{aligned} \mathrm{EU}_{\mathrm{CDT}}(T_1) &= \mathrm{u}(T_1 \,\&\, P_1) \cdot \mathrm{cr}(T_1 \mathbin{\square\!\!\rightarrow} P_1) + \mathrm{u}(T_1 \,\&\, P_2) \cdot \mathrm{cr}(T_1 \mathbin{\square\!\!\rightarrow} P_2) \\ &= 1,000,000 \cdot \mathrm{cr}(T_1 \mathbin{\square\!\!\rightarrow} P_1) + 0 \cdot \mathrm{cr}(T_1 \mathbin{\square\!\!\rightarrow} P_2) \end{aligned}$$

$$\begin{aligned} \mathrm{EU}_{\mathrm{CDT}}(T_2) &= \mathrm{u}(T_2 \,\&\, P_1) \cdot \mathrm{cr}(T_2 \mathbin{\square\!\!\rightarrow} P_1) + \mathrm{u}(T_2 \,\&\, P_2) \cdot \mathrm{cr}(T_2 \mathbin{\square\!\!\rightarrow} P_2) \\ &= 1,001,000 \cdot \mathrm{cr}(T_2 \mathbin{\square\!\!\rightarrow} P_1) + 1,000 \cdot \mathrm{cr}(T_2 \mathbin{\square\!\!\rightarrow} P_2) \end{aligned} \tag{7.18}$$

It doesn't matter what particular values the credences in these expressions take, because the act has no causal influence on the prediction. That is,

$$\mathrm{cr}(T_1 \mathbin{\square\!\!\rightarrow} P_1) = \mathrm{cr}(P_1) = \mathrm{cr}(T_2 \mathbin{\square\!\!\rightarrow} P_1) \tag{7.19}$$

and

$$\mathrm{cr}(T_1 \mathbin{\square\!\!\rightarrow} P_2) = \mathrm{cr}(P_2) = \mathrm{cr}(T_2 \mathbin{\square\!\!\rightarrow} P_2) \tag{7.20}$$

With these causal independencies in mind, you can tell by inspection of Equation (7.18) that $\mathrm{EU}_{\mathrm{CDT}}(T_2)$ will be greater than $\mathrm{EU}_{\mathrm{CDT}}(T_1)$, and Causal Decision Theory endorses two-boxing.

7.3.3 Responses and extensions

So is that it for Evidential Decision Theory? Philosophical debates rarely end cleanly, and Evidential Decision Theorists have made a number of responses to the Newcomb Problem.

First, one might respond that one-boxing is the rationally mandated act. Representing the two-boxers, David Lewis once wrote

> The one-boxers sometimes taunt us: if you're so smart, why ain'cha rich? They have their millions and we have our thousands, and they think this goes to show the error of our ways. They think we are not rich because we have irrationally chosen not to have our millions. (1981b, p. 377)

Lewis' worry is this: Suppose a one-boxer and a two-boxer each go through the Newcomb scenario many times. As a successful predictor, the being in the story will almost always predict that the one-boxer will one-box, and so place the $1,000,000 in the second box for him. Meanwhile, the two-boxer will almost always find the second box empty. The one-boxer will rack up millions of dollars, while the two-boxer will gain only thousands. Each agent has the goal of making as much money as possible, so one-boxing (and, by extension, EDT) seems to provide a better rational strategy for reaching one's goals than two-boxing (and CDT).

The Causal Decision Theorist's response (going at least as far back as (Gibbard and Harper 1978/1981)) is that some unfortunate situations reward agents monetarily for behaving irrationally, and the Newcomb Problem is one of them. The jury is still out on whether this response is convincing. In November 2009 the PhilPapers Survey polled over three thousand philosophers, and found that 31.4% of them accepted or leaned towards two-boxing in the Newcomb Problem, while 21.3% accepted or leaned towards one-boxing. (The remaining respondents were undecided or offered a different answer.) So it's unclear that EDT's embrace of one-boxing is a fatal defect. Meanwhile, there are other cases in which EDT seems to give the intuitively rational result while CDT does not (Egan 2007).

Jeffrey, on the other hand, was convinced that two-boxing is rationally required in the Newcomb Problem. So he defended Evidential Decision Theory in different ways. In the second edition of *The Logic of Decision* (1983), Jeffrey added a **ratifiability** condition to his EDT. The idea of ratifiability is that an act is rationally permissible only if the agent assigns it the highest expected utility conditional on the supposition that he chooses to perform it. Ratifiability avoids regret—if choosing to perform an act would make you

wish you'd done something else, then you shouldn't choose it. In the Newcomb Problem, supposing that you'll choose to one-box makes you confident that the being predicted one-boxing, and so makes you confident that the $1,000,000 is in the second box. So supposing that you'll choose to one-box makes two-boxing seem the better choice. One-boxing is unratifiable, and so can be rationally rejected.

We won't cover the technical details of ratifiability here, in part because Jeffrey ultimately abandoned that response. Jeffrey eventually (1993, 2004) agreed with other commentators that Newcomb's Problem isn't really a decision problem. Suppose that in the Newcomb Problem the agent assigns the credences we reported earlier because she takes the causal structure of her situation to be something like Figure 7.2. In that case, she will see her physical brain state as having such a strong influence on how many boxes she takes that whether she one-boxes or two-boxes will no longer seem a free choice. Jeffrey held that in order to make a genuine decision, an agent must see her choice as the cause of the act (and ultimately the outcome) produced. Read in this light, the Newcomb case seemed to involve too much causal influence on the agent's act from factors besides her choice. In the last sentences of his final work, Jeffrey wrote, "I now conclude that in Newcomb problems, 'One box or two?' is not a question about how to choose, but about what you are already set to do, willy-nilly. Newcomb problems are not decision problems." (2004, p. 113)

7.4 Exercises

Unless otherwise noted, you should assume when completing these exercises that credence distributions under discussion satisfy the probability axioms and Ratio Formula. You may also assume that whenever a conditional probability expression occurs, the needed proposition has nonzero unconditional credence so that conditional probabilities are well-defined.

Problem 7.1. When you play craps in a casino there are a number of different bets you can make at any time. Some of these are "proposition bets" on the outcome of the next roll of two fair dice. Below is a list of some proposition bets, and their payouts. (A payout of 4 to 1 means that if you put down $1 and win the bet, you keep your original $1 plus an additional $4. If you lose the bet, you lose your $1.)

Name of Bet	Wins when	Payout
Big red	Dice total 7	4 to 1
Any craps	Dice total 2, 3, or 12	7 to 1
Snake eyes	Dice total 2	30 to 1

Suppose you place a $1 bet on each of the three propositions listed above. Rank the three bets from highest expected dollar value to lowest.

Problem 7.2. (a) Suppose an agent is indifferent between two gambles with the following utility outcomes:

	P	$\sim P$
Gamble 1	x	y
Gamble 2	y	x

where P is a proposition about the state of the world, and x and y are utility values with $x \neq y$. Assuming this agent maximizues $\mathrm{EU}_{\mathrm{SAV}}$, what can you determine about the agent's $\mathrm{cr}(P)$?

(b) Suppose the same agent is also indifferent between these two gambles:

	P	$\sim P$
Gamble 3	a	z
Gamble 4	m	m

where P is the same proposition as before, $a = 100$, and $z = -100$. What can you determine about m?

Problem 7.3. You are confronted with a decision problem involving two possible states of the world (S and $\sim S$) and three available acts (A, B, and C). Assume you are using Jeffrey's decision theory to determine your preferences.

(a) Suppose that of the three S-outcomes, $B \,\&\, S$ does not have the highest utility for you. Also, of the three $\sim S$-outcomes, $B \,\&\, \sim S$ does not have the highest utility. Does it follow that you should not choose act B? Defend your answer.

(b) Suppose that of the S-outcomes, $B \,\&\, S$ has the *lowest* utility for you. Also, of the three $\sim S$-outcomes, $B \,\&\, \sim S$ has the *lowest* utility. Does it follow that you should not choose act B? Defend your answer.*

*This problem was inspired by a problem of Brian Weatherson's.

Problem 7.4. Prove that the Dominance Principle follows from Savage's expected utility theory. (Restrict your discussion to *finite* partitions of acts and states.)

Problem 7.5. Referring to the payoff tables for Allais' Paradox in Section 7.2.4, show that no assignment of values to u(\$0), u(\$1M), and u(\$5M) that makes $\mathrm{EU}_{\mathrm{EDT}}(A) > \mathrm{EU}_{\mathrm{EDT}}(B)$ will also make $\mathrm{EU}_{\mathrm{EDT}}(D) > \mathrm{EU}_{\mathrm{EDT}}(C)$. (You may assume that the agent assigns equal credence to each numbered ticket's being selected, and this holds regardless of which gamble is made.)

Problem 7.6. Having gotten a little agressive on a routine single to center field, you're now halfway between first base and second base. The throw from the center fielder is in midair, but given the angle you can't tell whether it's headed to first or second. You must decide whether to proceed to second base or run back to first. Being out has zero utility for you; being safe is better; and being safe at second has twice the utility of being safe at first. However this center fielder has a great track-record at predicting where runners will go—your credence in his throwing to second conditional on your going there is 90%, while your credence in his throwing to first conditional on your going to first is 80%. (Assume that if you and the throw go to the same base, you will certainly be out, but if you and the throw go to different bases you'll certainly be safe.)

(a) Of the two acts available (running to first or running to second), which should you prefer according to Evidential Decision Theory (that is, accoring to Jeffrey's decision theory)?

(b) Does the problem provide enough information to determine which act is preferred by Causal Decision Theory? If so, explain which act is preferred. If not, explain what further information would be required and how it could be used to determine a preference.

Problem 7.7. In the Newcomb Problem, do you think it's rational to take just one box or take both boxes? Explain your thinking.

7.5 Further reading

Introductions and Overviews

Martin Peterson (2009). *An Introduction to Decision Theory.* Cambridge Introductions to Philosophy. Cambridge: Cambridge University Press

A book-length general introduction to decision theory, including chapters on game theory and social choice theory.

Classic Texts

Leonard J. Savage (1954). *The Foundations of Statistics*. New York: Wiley

Savage's classic book laid the foundations for modern decision theory and much of contemporary Bayesian statistics.

Richard C. Jeffrey (1983). *The Logic of Decision*. 2nd. Chicago: University of Chicago Press

In the first edition, Jeffrey's Chapter 1 introduced a decision theory capable of handling dependent acts and states. In the second edition, Jeffrey added an extra section to this chapter explaining his "ratifiability" response to the Newcomb Problem.

Extended Discussion

Lara Buchak (2013). *Risk and Rationality*. Oxford: Oxford University Press

Presents a generalization of the decision theories discussed in this chapter that is consistent with a variety of real-life agents' responses to risk. For instance, Buchak's theory accommodates genuine risk-aversion, and allows agents to simultaneously prefer Gamble A to Gamble B and Gamble D to Gamble C in Allais' Paradox.

James M. Joyce (1999). *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press

A systematic explanation and presentation of causal decision theory, unifying that approach under a general framework with evidential decision theory and proving a representation theorem that covers both. (Note that Joyce introduces a special kind of

Notes

[1]The law of large numbers actually comes in many forms, each of which has slightly different conditions and a slightly different conclusion. Most versions require the repeated trials to be independent and identically distributed (IID), meaning that each trial has the

same probability of yielding a given result and the result on a given trial is independent of all previous results. (In other words, you think your batter is consistent across games and unaffected by previous performance.) Most versions also assume Countable Additivity for their proof. Finally, since we are dealing with results involving the infinite, we should remember that in such cases credence 1 doesn't necessarily mean certainty. An agent who satisfies the probability axioms, the Ratio Formula, and Countable Additivity will assign credence 1 to the average's approaching the expectation in the limit, but that doesn't mean she *rules out* all possibilities in which those values don't converge. (For Countable Additivity and cases of credence-1 that don't mean certainty, see Section 5.4. For more details and proofs concerning laws of large numbers, see (Feller 1968, Ch. X).)

[2]See (Bernoulli 1738/1954) for both his discussion and a reference to Cramer.

[3]Although Savage didn't actually approach things this way, to simplify presentation I will treat acts, states, and outcomes as propositions—the proposition that the agent will perform the act, the proposition that the world is in a particular state, and the proposition that a particular outcome occurs.

[4]The Dominance Principle I've presented is sometimes known as the Strong Dominance Principle. The Weak Dominance Principle says that if A produces *at least as good* an outcome as B in each possible state of the world, plus a better outcome in at least one possible state of the world, then A is preferred to B. Weak Dominance is also a consequence of Savage's expected utility theory, and has the same problems as Strong Dominance.

[5]In a similar display of poor reasoning, Shakespeare's Henry V (Act 4, Scene 3) responds to Westmoreland's wish for more troops on their side of the battle—"O that we now had here but one ten thousand of those men in England, that do no work today"—with the following:

> If we are marked to die, we are enough to do our country loss;
> and if to live, the fewer men, the greater share of honor.
> God's will, I pray thee wish not one man more.

[6]For a brief discussion and references, see (Jeffrey 1983, §1.8).

[7]Instead of referring to "acts", "states", "outcomes", and "utilities", Jeffrey speaks of "acts", "conditions", "consequences", and "desirabilities" (respectively). As in my presentation of Savage's theory, I have made some changes to Jeffrey's approach for the sake of simplicity and consistency with the rest of the discussion.

[8]The decision-theoretic structure here bears striking similarities to Simpson's Paradox. We saw in Section 3.2.3 that while David Justice had a better batting average than Derek Jeter in each of the years 1995 and 1996, over the entire two-year span Jeter's average was better. This was because Jeter had a much higher proportion of his bats in 1996, which was a better year for both hitters. So selecting a Jeter at-bat is much more likely to land you in a good year for hitting. Similarly, the deterrence utility table shows that disarming yields better outcomes than arming on each possible state of the world. Yet arming is much more likely than disarming to land you in the peace state (the right-hand column of the table), and so get you a desirable outcome.

[9]While Savage coined the phrase "Sure-Thing Principle", it's actually a bit difficult to tell from his text exactly what he meant by it. I've presented a contemporary cleaning-up of Savage's discussion, inspired by the Sure-Thing formulation in (Eells 1982, p. 10). It's also worth noting that the Sure-Thing Principle is intimately related to decision-theoretic axioms known as Separability and Independence, but we won't delve into those conditions here.

[10]By the way, in case you're looking for a clever way out Nozick specifies in a footnote

to the problem that if the being predicts you will decide what to do via some random process (like flipping a coin), he does not put the $1,000,000 in the second box.

[11]It's important for Causal Decision Theory that $A \mathrel{\Box\!\!\rightarrow} S$ conditionals be "causal" counterfactuals rather than "backtracking" counterfactuals; we hold facts about the past fixed when assessing A's influence on S. (See (Lewis 1981a) for the distinction and some explanation.)

[12]There are actually many ways of executing a causal decision theory; the approach presented here is that of (Gibbard and Harper 1978/1981), drawing from (Stalnaker 1972/1981). Lewis (1981a) thought Causal Decision Theory should instead return to Savage's unconditional credences and independence assumptions, but with the specification that acts and states be *causally* independent. For a comparison of these approaches along with various others, plus a general formulation of Causal Decision Theory that attempts to cover them all, see (Joyce 1999).