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 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 *The Foundations of Statistics*, Leonard 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. (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 Reading section of this chapter.) As a result, I haven't packed as much information into this chapter as into the preceding chapter on confirmation. I hope only to equip the reader with the terminology and ideas we will need later in this book, and that she would need to delve further into the philosophy of decision theory.

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 Jeffrey's Evidential Decision Theory, which improves on Savage's 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 proposed as

Fundamentals of Bayesian Epistemology 2: Arguments, Challenges, Alternatives. Michael G. Titelbaum, Oxford University Press. © Michael G. Titelbaum 2022. DOI: 10.1093/oso/9780192863140.003.0007

a better response to Newcomb. I will close 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—say, the number of hits a particular batter will have in tonight's baseball game—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 one hit, 20% confident she'll have two hits, and 50% confident she'll have three, your expectation for the number of hits is

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

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 a kind of *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.

While your expectation for a quantity isn't necessarily the exact value you think it will take on a given occasion, it should equal the *average* value you expect that quantity to take in the long run. Suppose you're certain that our batter will play in many, many games. 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, her average number of hits per game will tend toward 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]

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. If you buy shares for less than $46 each, your expectation for that transaction will be positive (you'll expect to make money on the deal). If you buy shares for more than $46, you'll expect to lose money.

The idea that your fair price for an investment should equal your expectation of its monetary return dates to Blaise Pascal, in a famous seventeenth-century correspondence with Pierre Fermat (Fermat and Pascal 1654/1929). There are a couple of reasons why this is a sensible idea. 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 you should anticipate a long-run 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, each share becomes more valuable to you, and you should be willing to pay a higher price. This is exactly what the expected value calculation predicts. If you learn that a good earnings report will send the share value to only $50, this decreases the expected value of the investment and also decreases the price you should 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.

We can also think about fair betting prices using odds. We saw in Section 2.3.4 that an agent's odds against P equal $\text{cr}(\sim P) : \text{cr}(P)$. So if the agent's credence in P is 0.25, her odds against P are 3 : 1. What will she consider a fair bet on P? Consider what the casinos would call a bet on P at 3 : 1 odds. If you place such a bet and win, you get back the original amount you bet plus three times that amount. If you lose your bet, you're out however much you bet. In terms of net returns, a bet at 3 : 1 odds offers you a possible net gain that's three times your possible net loss.

So suppose an agent with 0.25 credence in P places a \$20 bet on P at 3 : 1 odds. Her expected net return is

$$\begin{aligned}&(\text{net return on winning bet}) \cdot \text{cr}(P) + (\text{net return on losing bet}) \cdot \text{cr}(\sim P)\\ &\qquad = \$60 \cdot 0.25 + -\$20 \cdot 0.75 = \$0 \qquad (7.4)\end{aligned}$$

This agent expects a bet on P at 3 : 1 odds to be a break-even gamble—from her perspective, it's a fair bet. She will be willing to bet on P at those odds or anything higher. In general, an agent who bets according to her expectations will accept a bet on a proposition at odds equal to her odds against it, or anything higher. Remember that an agent's odds against a proposition *increase* as her credence in the proposition *decreases*. So if an agent becomes less confident in P, you need to offer her higher odds on P before she'll be willing to gamble.

A lottery ticket is a type of bet, and in the right situation calculating its expected value can be highly lucrative. Ellenberg (2014, Ch. 11) relates the story of Massachusetts's Cash WinFall state lottery game, which was structured in such a way that if the jackpot got high enough, the expected payout for a single ticket grew larger 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 *extremely* effective.

7.1.1 The move to utility

Yet sometimes we value something 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 to ride, 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 with the notion of **utility**. Introduced by Daniel Bernoulli and Gabriel Cramer in the eighteenth 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 real-valued 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.

Utilities provide a uniform value-measurement scale. 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 reflects the values on which you truly make your decisions. Let's say that having no money is worth

0 utils to you in this case, receiving one dollar and being able to get on the bus is worth 100 utils, and receiving \$2.02 is worth 102 utils. (The larger amount of money is still more valuable to you; just not *much* more valuable.) 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 each additional dollar (or peso, or yuan, or...), the more dollars she already has the less extra utility it will be. The first billion you earn makes your family comfortable; the second billion doesn't have as much significance for your life. Postulating an underlying locus of value distinguishable from monetary 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 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 nuanced 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 rankings and money pumps

A **decision problem** presents an agent with a partition of **acts**, from which she must choose exactly one. Decision theory aims to lay down rational principles governing choices in decision problems. It does so by supposing that a rational agent's choice of acts tracks her preferences among those acts. If the available acts are A and B, and she prefers A to B (we write $A \succ B$), then the agent decides to perform action A. A similar point applies when $B \succ A$. Yet it might be that the agent is indifferent between A and B (we write $A \sim B$), in which case she is rationally permitted to choose either one.

Sometimes a decision among acts is easy. 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 uncertainty into her decision. She needs a technique for combining credences and utilities to generate preferences.

Decision theory responds to this problem by providing a **valuation function**, which combines credences and utilities to assign each act a numerical score. The agent's preferences are assumed to match these scores: $A \succ B$ just in case A receives a higher score than B, while $A \sim B$ when the scores are equal. Given a particular decision problem, a rational agent will select the available act with the highest score (or—if there are ties at the top—one of the acts with the highest score).

Here's an example of a valuation function, just to convey the idea: Suppose you assign each act a numerical score by considering all the possible worlds to which you assign nonzero credence, finding the one in which that act produces the lowest utility, and then assigning that minimal utility value as the act's score. This valuation function generates preferences satisfying the **maximin rule**, so called because it selects the act with the highest minimum utility payoff. Maximin attends to only the worst case scenario for each available act.

While maximin is just one valuation function (we'll see others later), any approach that ties preferences to numerical scores assigned over acts imposes a certain structure on an agent's preferences. For instance, it guarantees that her preferences will display:

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.

This follows from the simple fact that numerical inequalities are transitive: each act's score is a number, so if act A's score is greater than act B's, and B's is greater than C's, then A's must be greater than C's as well.

Preference Transitivity will be endorsed as a rational constraint by any decision theory that ties preferences to numerical valuation functions. One might object 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 such an agent's preference ranking could go silent on the comparison between A and C and still be rational. Yet by coordinating preference with a numerical valuation over the entire partition of acts, we have already settled this issue; we have required the agent's preferences to form a complete ranking. Since every act receives a score, every act is comparable, and our theory demands the agent assign a preference (or indifference) between any two acts. 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.

Notice that Preference Completeness entails the following:

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

To recap: Decision theory begins by requiring an agents' choices to reflect her preferences, then coordinates those preferences with a numerical valuation function combining credences and utilities. By making the latter move, decision theory requires preferences to satisfy Preference Transitivity and Asymmetry. Hopefully it's intuitive that rational preferences satisfy these two conditions. But we can do better than that: We can provide an *argument* that Preference Transivity and Asymmetry are rational requirements.

Consider a situation in which some of us find ourselves frequently. 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 added 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 cooking for himself to eating at a 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.[3]

Violating Preference Transitivity leaves one susceptible to a money-pumping set of trades. (If you violate Preference Asymmetry, the money pump is even simpler.) In a money pump, the agent proceeds through a series of exchanges, each of which looks favorable given his preferences between the two acts 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 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.[4]

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.[5]

Here's a simple example: Suppose you're trying to decide whether to carry an umbrella today, but you're uncertain whether it's going to rain. 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 outcomes produced by various act/state combinations. 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).

How should you evaluate available acts and set your preferences among them? For a finite partition $\{S_1, S_2, \dots, S_n\}$ of possible states of the world, Savage offers the following valuation function:

$$\begin{aligned} \mathrm{EU}_{\mathrm{SAV}}(A) = \mathrm{u}(A \mathbin{\&} S_1) \cdot \mathrm{cr}(S_1) + \mathrm{u}(A \mathbin{\&} S_2) \cdot \mathrm{cr}(S_2) \\ + \dots + \mathrm{u}(A \mathbin{\&} S_n) \cdot \mathrm{cr}(S_n) \end{aligned} \tag{7.5}$$

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; $\mathrm{u}(A \mathbin{\&} S_i)$ is the utility she assigns to the outcome that will eventuate should she perform act A in state S_i.[6] 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 decision among a partition of acts, a rational agent will set her preferences in line with her expected utilities. She will choose to perform an act with at least as great an expected utility as that of any act on offer.

Now suppose that in the umbrella case you have a 0.30 credence in rain. We can calculate expected utilities for each of the available acts as follows:

$$\begin{aligned} \mathrm{EU}_{\mathrm{SAV}}(\text{take}) &= 0 \cdot 0.30 + -1 \cdot 0.70 = -0.7 \\ \mathrm{EU}_{\mathrm{SAV}}(\text{leave}) &= -10 \cdot 0.30 + 0 \cdot 0.70 = -3 \end{aligned} \tag{7.6}$$

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 ranking satisfying Preference 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[7] seems intuitively like a good rational principle. Yet, surprisingly, there are decision problems in which it yields very bad results. 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 don't 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.*[8] The Dominance Principle—and Savage's expected utility theory in general—breaks down when the state of the world is influenced by which 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.[9] 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,[10] Jeffrey calculates

$$\begin{aligned} \mathrm{EU_{EDT}}(A) = \mathrm{u}(A \mathbin{\&} S_1) \cdot \mathrm{cr}(S_1 \mid A) + \mathrm{u}(A \mathbin{\&} S_2) \cdot \mathrm{cr}(S_2 \mid A) \\ + \ldots + \mathrm{u}(A \mathbin{\&} S_n) \cdot \mathrm{cr}(S_n \mid A) \end{aligned} \tag{7.7}$$

I'll explain the "EDT" subscript later on; for now, it's crucial to see that Jeffrey alters Savage's approach (Equation (7.5)) 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 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.

At a typical dinner party, the entree for the evening is settled well before the guests arrive. But let's suppose that tonight's host is especially accommodating, and 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.) The guest is 75% confident that the host will select the meat that best pairs with the wine provided. Thus 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 together require all the credences conditional on white wine sum to 1, so the values in the first row sum to 1. The values in the second row sum to 1 for a similar reason. (In this example the values in each column sum to 1 as well, but that won't always be the case.)

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

$$\begin{aligned} \mathrm{EU}_{\mathrm{EDT}}(\text{white}) &= \mathrm{u}(\text{white}\ \&\ \text{chicken}) \cdot \mathrm{cr}(\text{chicken} \mid \text{white}) \\ &\qquad + \mathrm{u}(\text{white}\ \&\ \text{beef}) \cdot \mathrm{cr}(\text{beef} \mid \text{white}) \\ &= 1 \cdot 0.75 + -1 \cdot 0.25 \\ &= 0.5 \end{aligned} \tag{7.8}$$

(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 distribution. That is, an act A and state S_i are independent for Jeffrey just in case

$$\mathrm{cr}(S_i \mid A) = \mathrm{cr}(S_i) \tag{7.9}$$

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 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 the advocate of nuclear deterrence takes 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:

$$\begin{aligned} \mathrm{EU}_{\mathrm{EDT}}(\mathrm{arm}) &= -100 \cdot 0.1 + 0 \cdot 0.9 = -10 \\ \mathrm{EU}_{\mathrm{EDT}}(\mathrm{disarm}) &= -50 \cdot 0.8 + 50 \cdot 0.2 = -30 \end{aligned} \tag{7.10}$$

Relative to the deterrence advocate's credences, Jeffrey's theory yields a preference for arming. Act/state dependence has created a preference ranking at odds with the Dominance Principle.[11] 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 is reliable. 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

Different people respond to risks differently. 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. On this explanation, the agent would be happy to take a risk if only it offered her a higher expectation of what she really values: utility. But might some agents be *genuinely* risk-averse—might they be 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 simple risk-aversion cases involving monetary gambles can be explained by attributing to the agent a utility distribution with decreasing marginal utility 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 get to choose between two gambles, 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

(Here "\$1M" is short for 1 million dollars.) 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[12]

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 therefore also 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 those 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 7.10.)

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 this combination of preferences could be perfectly rational. 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.[13]

We have been discussing 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 many real-life subjects 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 or 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 distribution 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. 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 focus on 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. 114–15)

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 act.[14] 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 whether 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. I reason through my decision as follows: Drinking a third appletini is the kind of act highly 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 don't order that next appletini.

Let's work through the reasoning just described using decision theory. First, stipulate that I have the following utility table:

	smoker	non
third appletini	−99	1
stop at two	−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):

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

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

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:

$$\mathrm{cr}(S \mid A) > \mathrm{cr}(S \mid {\sim}A) \tag{7.13}$$

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. Luckily, Jeffrey's decision theory works even when acts and states are dependent. So I apply Jeffrey's valuation function to calculate expected utilities for the two acts:

$$\begin{aligned} \mathrm{EU}_{\mathrm{EDT}}(A) &= -99 \cdot \mathrm{cr}(S \mid A) + 1 \cdot \mathrm{cr}({\sim}S \mid A) \\ \mathrm{EU}_{\mathrm{EDT}}({\sim}A) &= -100 \cdot \mathrm{cr}(S \mid {\sim}A) + 0 \cdot \mathrm{cr}({\sim}S \mid {\sim}A) \end{aligned} \tag{7.14}$$

Looking at these equations, you might think that A receives the higher expected utility. But I assign a considerably higher value to $\text{cr}(S \mid A)$ than $\text{cr}(S \mid {\sim}A)$, 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 higher expected utility and I prefer to perform ${\sim}A$.

But this reasoning 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 any of that. So I might as well enjoy the drink.[15]

Assuming the reasoning in the previous paragraph—rather than the reasoning originally presented in the example—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 (7.12)). Having an addictive personality is unfortunate because of its potential consequences for becoming a smoker. I expect the world in which I order another drink to be a worse world than the world in which I don't, and this is reflected in the EU_{EDT} calculation. Jeffrey's theory assesses the act of ordering a third appletini not in terms of outcomes it will *cause* to come about, but instead in terms of outcomes it provides *evidence* for. 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 an outcome that it's too late for her to *cause* (or *prevent*). Even though the act indicates the outcome, it seems irrational to factor the value of that outcome 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 avoid becoming a smoker would be "a futile attempt to manipulate the cause by suppressing its symptoms." In making decisions we should focus on what we can control—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)

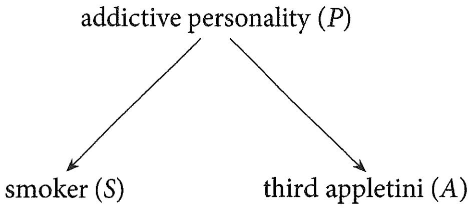

Figure 7.1 Third drink causal fork

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 correlated even though neither causes the other. This accounts for A's being relevant to S on my credence distribution (Equation (7.13)) even though my ordering the third appletini has no causal influence on whether I'll become a smoker.

The causally spurious correlation in my credences affects Jeffrey's expected utility calculation because that calculation works with credences in states conditional on acts ($\text{cr}(S_i \mid A)$). Jeffrey replaced Savage's $\text{cr}(S_i)$ with this conditional expression to track dependencies between states and acts. The Causal Decision Theorist responds that while credal correlation is a *kind* of dependence, it's not the kind of dependence that decisions should track. Preferences should be based on *causal* dependencies. So the Causal Decision Theorist's valuation function is:

$$\begin{aligned} \text{EU}_{\text{CDT}}(A) = {} & \text{u}(A \,\&\, S_1) \cdot \text{cr}(A \mathrel{\Box\!\!\rightarrow} S_1) + \text{u}(A \,\&\, S_2) \cdot \text{cr}(A \mathrel{\Box\!\!\rightarrow} S_2) \\ & + \ldots + \text{u}(A \,\&\, S_n) \cdot \text{cr}(A \mathrel{\Box\!\!\rightarrow} S_n) \end{aligned} \tag{7.15}$$

Here $A \mathrel{\Box\!\!\rightarrow} S$ represents the subjunctive conditional "If the agent were to perform act A, state S would occur."[16] Causal Decision Theory uses such conditionals to track causal relations in the world.[17] Of course, an agent may be uncertain what consequences a given act A would cause. So EU_{CDT} looks across the partition $\{S_1, \ldots, S_n\}$, and invokes the agent's credences that A would cause various states S_i to occur.

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

$$\text{cr}(\text{chicken} \mid \text{white}) = \text{cr}(\text{white} \mathrel{\Box\!\!\rightarrow} \text{chicken}) = 0.75 \tag{7.16}$$

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 involving 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 distribution—that is, when $\mathrm{cr}(S \mid A) = \mathrm{cr}(S)$. CDT focuses on whether the agent takes A and S to be *causally* independent, which occurs just when

$$\mathrm{cr}(A \,\Box\!\!\rightarrow S) = \mathrm{cr}(S) \tag{7.17}$$

When an agent thinks A has no causal influence on S, her 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 I know 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 CDT would have me prefer the dominant act and order the third appletini.

Now we can return to the Newcomb Problem, focusing on a version of it 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:

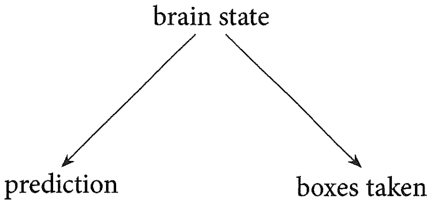

Figure 7.2 Newcomb Problem causal fork

$$\begin{aligned} \text{EU}_{\text{EDT}}(T_1) &= \text{u}(T_1 \,\&\, P_1) \cdot \text{cr}(P_1 \mid T_1) + \text{u}(T_1 \,\&\, P_2) \cdot \text{cr}(P_2 \mid T_1) = 990{,}000 \\ \text{EU}_{\text{EDT}}(T_2) &= \text{u}(T_2 \,\&\, P_1) \cdot \text{cr}(P_1 \mid T_2) + \text{u}(T_2 \,\&\, P_2) \cdot \text{cr}(P_2 \mid T_2) = 11{,}000 \end{aligned} \tag{7.18}$$

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; 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:

$$\begin{aligned} \text{EU}_{\text{CDT}}(T_1) &= \text{u}(T_1 \,\&\, P_1) \cdot \text{cr}(T_1 \mathrel{\Box\!\!\rightarrow} P_1) + \text{u}(T_1 \,\&\, P_2) \cdot \text{cr}(T_1 \mathrel{\Box\!\!\rightarrow} P_2) \\ &= 1{,}000{,}000 \cdot \text{cr}(T_1 \mathrel{\Box\!\!\rightarrow} P_1) + 0 \cdot \text{cr}(T_1 \mathrel{\Box\!\!\rightarrow} P_2) \\ \\ \text{EU}_{\text{CDT}}(T_2) &= \text{u}(T_2 \,\&\, P_1) \cdot \text{cr}(T_2 \mathrel{\Box\!\!\rightarrow} P_1) + \text{u}(T_2 \,\&\, P_2) \cdot \text{cr}(T_2 \mathrel{\Box\!\!\rightarrow} P_2) \\ &= 1{,}001{,}000 \cdot \text{cr}(T_2 \mathrel{\Box\!\!\rightarrow} P_1) + 1{,}000 \cdot \text{cr}(T_2 \mathrel{\Box\!\!\rightarrow} P_2) \end{aligned} \tag{7.19}$$

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,

$$\text{cr}(T_1 \mathrel{\Box\!\!\rightarrow} P_1) = \text{cr}(P_1) = \text{cr}(T_2 \mathrel{\Box\!\!\rightarrow} P_1) \tag{7.20}$$

and

$$\text{cr}(T_1 \mathrel{\Box\!\!\rightarrow} P_2) = \text{cr}(P_2) = \text{cr}(T_2 \mathrel{\Box\!\!\rightarrow} P_2) \tag{7.21}$$

With these causal independencies in mind, you can tell by inspection of Equation (7.19) that $\text{EU}_{\text{CDT}}(T_2)$ will be greater than $\text{EU}_{\text{CDT}}(T_1)$, and Causal Decision Theory endorses two-boxing.[18]

7.3.3 Responses and extensions

So is that it for Evidential Decision Theory? Philosophical debates rarely end cleanly; 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's worry is this: Suppose a one-boxer and a two-boxer each go through the Newcomb scenario many times. As a highly accurate 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 toward two-boxing in the Newcomb Problem, while 21.3% accepted or leaned toward one-boxing. (The remaining respondents were undecided or offered a different answer.) So not everyone considers EDT's embrace of one-boxing 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 tried to reconcile Evidential Decision Theory with that verdict in a variety of ways. In the second edition of *The Logic of Decision* (1983), Jeffrey added a **ratifiability** condition to his EDT. Ratifiability holds that an act is rationally permissible only if the agent assigns it the highest expected utility conditional on the supposition that she 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) came to believe that the Newcomb Problem isn't really a decision problem, and therefore isn't the kind of thing against which a decision theory (like EDT) should be tested. Suppose that in the Newcomb Problem the agent assigns the credences we described 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 like 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 beyond her choice. In the final sentences of his last 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 the credence distributions under discussion satisfy the probability axioms and Ratio Formula. You may also assume that whenever a conditional credence 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 the odds at which casinos offer them:

Name of bet	Wins when	Odds paid
Big red	Dice total 7	4 : 1
Any craps	Dice total 2, 3, or 12	7 : 1
Snake eyes	Dice total 2	30 : 1

Suppose you place a $1 bet on each proposition at the odds listed above. Rank the three bets from highest expected net return to lowest.

Problem 7.2. 🌶 Suppose you're guarding Stephen Curry in an NBA game and he is about to attempt a three-point shot. You have to decide whether to foul him in the act of shooting.

(a) If you don't foul him, he will attempt the shot. During the 2014–15 NBA season, Steph Curry made 44.3% of his three-point shot attempts. What is the expected number of points you will yield on Curry's shot attempt if you decide not to foul him?

(b) Suppose that if you decide to foul Curry, you can ensure he doesn't get a three-point shot attempt off. However, your foul will send him to the free-throw line, where he will get three attempts, each worth one point if he makes it. During the 2014–15 NBA season, Curry made 91.4% of his free-throw attempts. Assuming that the result of each free-throw attempt is probabilistically independent of the results of all the others, what is the expected number of points you will yield on Curry's free throws if you foul him?

(c) Given your calculations from parts (a) and (b), should you foul Steph Curry when he attempts a three-pointer?

Problem 7.3. The St. Petersburg game is played as follows: A fair coin is flipped repeatedly until it comes up heads. If the coin comes up heads on the first toss, the player wins $2. Heads on the second toss pays $4, heads on the third toss pays $8, etc.[19]

(a) 🌶 If you assign fair prices equal to expected monetary payouts (and credences equal to objective chances), how much should you be willing to pay to play the St. Petersburg game?

(b) 🪶 If you were confronted with this game in real life, how much would you be willing to pay to play it? Explain your answer.

Problem 7.4. 🌶 Asked to justify his decision to bring along his chicken-replace-inator, Dr. Doofenshmirtz replies: "I'd rather have it and not need it than need it and not have it."

(a) Supposing the relevant states of the world are N (need-replace-inator) and $\sim N$, and the relevant acts are B (bring-replace-inator) and $\sim B$, what two outcomes is Doofenshmirtz referencing, and how is he claiming their utilities compare for him?

(b) Explain why on Savage's utility theory, this fact about Doofenshmirtz's utilities does not necessarily make his decision rationally permissible.

Problem 7.5. 🌶 Consider once again the utility table for the umbrella decision problem on page 255. Given this utility distribution, how confident would you need to be in rain for Savage's decision theory to recommend that you take your umbrella?

Problem 7.6. 🌶 Imagine there's some proposition P in which I'm highly interested (and whose truth I view as probabilistically independent of my behavior). Learning of my interest, a nefarious character offers to sell me the following betting ticket for \$0.70:

> This ticket entitles the bearer
> to \$1 if P is true,
> and nothing otherwise.

(a) For my fair betting price for this ticket to be exactly \$0.70, what would my credence in P have to be?

(b) The nefarious character also has a second ticket available, which he offers to sell me for \$0.70 as well:

> This ticket entitles the bearer
> to \$1 if $\sim P$ is true,
> and nothing otherwise.

For my fair betting price in this second ticket to be exactly \$0.70, what would my credence in $\sim P$ have to be?

(c) Suppose I throw caution to the wind and purchase both tickets, each at a price of \$0.70. Without knowing my actual credences in P and $\sim P$, can you nevertheless calculate my expected *total* monetary value for the two tickets combined—taking into account both what I spent to get them and what they might pay out?

Problem 7.7. 🌶🌶

(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 maximizes $\mathrm{EU}_{\mathrm{SAV}}$, what can you determine about the agent's cr(P)?

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

	P	$\sim P$
Gamble 3	d	w
Gamble 4	m	m

where $\mathrm{cr}(P) = \mathrm{cr}(\sim P)$, $d = 100$, and $w = -100$. What can you determine about m?

(c) Finally, suppose the agent is indifferent between these two gambles:

	Q	$\sim Q$
Gamble 5	r	s
Gamble 6	t	t

where $r = 100$, $s = 20$, and $t = 80$. What can you determine about cr(Q)?

Problem 7.8. 🌶🌶 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).

(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. Applying Savage's decision theory, 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. Still applying Savage's decision theory, does it follow that you should not choose act B? Defend your answer.

(c) Suppose now that you apply Jeffrey's decision theory to the situation in part (b). Do the same conclusions necessarily follow about whether you should choose act B? Explain.[20]

Problem 7.9. 🌶🌶 Suppose an agent faces a decision problem with two acts A and B and finitely many states.

(a) Prove that if the agent sets her preferences using $\mathrm{EU}_{\mathrm{SAV}}$, those preferences will satisfy the Dominance Principle.

(b) If the agent switches from $\mathrm{EU}_{\mathrm{SAV}}$ to $\mathrm{EU}_{\mathrm{EDT}}$, exactly where will your proof from part (a) break down?

Problem 7.10. 🌶🌶 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 7.11. 🌶🌶 Having gotten a little aggressive on a routine single to center field, you're now halfway between first base and second base. You must decide whether to proceed to second base or run back to first.

The throw from the center fielder is in midair, and given the angle you can't tell whether it's headed to first or second base. But you do know that 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%.

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. Being out has the same utility for you no matter where you're out. Being safe at first is better than being out, and being safe at second is better than being safe at first by the same amount that being safe at first is better than being out.

(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 7.12. ✎ 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

Michael D. Resnik (1987). *Choices: An Introduction to Decision Theory.* Minneapolis: University of Minnesota Press

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

Each of these provides 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 edition. 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.

Notes

1. The law of large numbers comes in many different 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 our 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 this context 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. The first money pump was presented by Davidson, McKinsey, and Suppes (1955, p. 146), who attributed the inspiration for their example to Norman Dalkey. I don't know who introduced the "money pump" terminology.

   By the way, if you've ever read Dr. Seuss's story "The Sneetches", the Fix-it-Up Chappie (Sylvester McMonkey McBean) gets a pretty good money pump going before he packs up and leaves.
4. Though Quinn (1990) presents a case ("the puzzle of the self-torturer") in which it may be rational for an agent to have intransitive preferences.
5. While Savage thought of acts as functions from states to outcomes, it will be simpler for us to 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.
6. For simplicity's sake we set aside cases in which some S_i make particular acts impossible. Thus A & S_i will never be a contradiction.
7. 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*. The names of the principles can be a bit confusing—it's not that Strong Dominance is a stronger *principle*; it's that it involves a stronger kind of dominance. In fact, the Weak Dominance Principle is logically stronger than the Strong Dominance Principle, in the sense that the Weak Dominance Principle entails the Strong Dominance Principle. (Thanks to David Makinson for suggesting this clarification.)

Despite being a logically stronger principle, Weak Dominance is also a consequence of Savage's expected utility theory, and has the same kinds of problems as Strong Dominance.

8. 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.

9. For a brief discussion and references, see Jeffrey (1983, § 1.8).
10. 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.
11. The decision-theoretic structure here bears striking similarities to Simpson's Paradox. We saw in Section 3.2.3 that while DeMar DeRozan had a better overall field-goal percentage than James Harden during the 2016–17 NBA season, from each distance (two-pointer versus three-pointer) Harden was more accurate. This was because a much higher proportion of DeRozan's shot attempts were two-pointers, which are much easier to make. So if you selected a DeRozan attempt at random, it was much more likely than a Harden attempt to have been a two-pointer, and so much more likely to have been made. 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.
12. 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 principles known as Separability and Independence, but we won't delve into those here.
13. Heukelom (2015) provides an accessible history of the Allais Paradox, and of Allais' disputes with Savage over it. Another well-known counterexample to Savage's decision theory based on risk aversion is the Ellsberg Paradox, which we'll discuss in Section 14.1.3.

14. In case you're looking for a clever way out of Newcomb's Problem, Nozick specifies in a footnote 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.
15. Eells (1982, p. 91) gives a parallel example from theology: "Calvinism is sometimes thought to involve the thesis that election for salvation and a virtuous life are effects of a common cause: a certain kind of soul. Thus, while leading a virtuous life does not cause one to be elected, still the probability of salvation is higher conditional on a virtuous life than conditional on an unvirtuous life. Should one lead a virtuous life?"
16. 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.)
17. 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).
18. If you feel like Newcomb's Problem is too fanciful and our appletini example too frivolous to merit serious concern, consider that Gallo et al. (2018) found substantial evidence that smoking more cigarettes or smoking for a longer time is correlated with a decreased risk of developing Parkinson's disease. If Reichenbach's Principle of the Common Cause (Section 3.2.4) is true, then either smoking has a causal effect on whether one develops Parkinson's, Parkinson's somehow affects whether one smokes, or some other cause makes one both more likely to smoke and less likely to develop Parkinson's. Pursuing this third, causal-fork option, the researchers speculated that a dopamine shortage in the brain might contribute both to Parkinson's and to a "low-risk-taking personality trait" that makes people less likely to smoke or more likely to quit. If that's right, then should you take up smoking to avoid Parkinson's? Other studies have found that high levels of education correlate positively with developing Parkinson's (Frigerio et al. 2005). Should you cut short your education to avoid the disease?
19. This game was invented by Nicolas Bernoulli in the eighteenth century.
20. This problem was inspired by a problem of Brian Weatherson's.

PART IV
ARGUMENTS FOR BAYESIANISM

To my mind, the best argument for Bayesian epistemology is the uses to which it can be put. In the previous part of this book we saw how the Bayesian approach interacts with confirmation and decision theory, two central topics in the study of theoretical and practical rationality (respectively). The five core normative Bayesian rules grounded formal representations of how an agent should assess what her evidence supports and how she should make decisions in the face of uncertainty. These are just two of the many applications of Bayesian epistemology, which have established its significance in the minds of contemporary philosophers.

Nevertheless, Bayesian history also offers more direct arguments for the normative Bayesian rules. The idea is to *prove* from premises plausible on independent grounds that, say, a rational agent's unconditional credences satisfy Kolmogorov's probability axioms. The three most prominent kinds of arguments for Bayesianism are those based on representation theorems, Dutch Books, and accuracy measurements. This part of the book will devote one chapter to each type of argument.

Some of these argument-types can be used to establish more than just the probability axioms as requirements of rationality; the Ratio Formula, Conditionalization, Countable Additivity, and other norms we have discussed may be argued for. Each argument-type has particular norms it can and can't be used to support; I'll mention these applications as we go along. But they *all* can be used to argue for probabilism.

As I mentioned in Chapter 2, probabilism is the thesis that a rational agent's unconditional credence distribution at a given time satisfies Kolmogorov's three axioms. (I sometimes call a distribution that satisfies the axioms a "probabilistic" distribution.) Among the probability axioms, by far the most difficult to establish as a rational requirement is Finite Additivity. We'll see why as we dig into the arguments' particulars, but it's worth a quick reminder at this point what Finite Additivity does.

Chapter 2 introduced three characters: Mr. Prob, Mr. Weak, and Mr. Bold. We imagine there is some single proposition P for which the three of them assign the following credences:

Mr. Prob:	$\mathrm{cr}(\mathsf{F}) = 0$	$\mathrm{cr}(P) = 1/6$	$\mathrm{cr}({\sim}P) = 5/6$	$\mathrm{cr}(\mathsf{T}) = 1$
Mr. Weak:	$\mathrm{cr}(\mathsf{F}) = 0$	$\mathrm{cr}(P) = 1/36$	$\mathrm{cr}({\sim}P) = 25/36$	$\mathrm{cr}(\mathsf{T}) = 1$
Mr. Bold:	$\mathrm{cr}(\mathsf{F}) = 0$	$\mathrm{cr}(P) = 1/\sqrt{6}$	$\mathrm{cr}({\sim}P) = \sqrt{5}/\sqrt{6}$	$\mathrm{cr}(\mathsf{T}) = 1$

All three of these characters satisfy the Non-Negativity and Normality axioms. They also satisfy such intuitive credence norms as Entailment—the rule that a proposition must receive at least as much credence as any proposition that entails it. We could easily introduce conditional credences that have them satisfy the Ratio Formula as well. Yet of the three, only Mr. Prob satisfies Finite Additivity. This demonstrates that Finite Additivity is logically independent of these other norms; they can be satisfied even if Finite Additivity is not.

Mr. Weak's credences are obtained by squaring each of Mr. Prob's. This makes Mr. Weak's levels of confidence in logically contingent propositions (P, ${\sim}P$) lower than Mr. Prob's. Mr. Weak is comparatively conservative, unwilling to be very confident in contingent claims. So while Mr. Weak is certain of $P \vee {\sim}P$, his individual credences in P and ${\sim}P$ sum to less than 1. Mr. Bold's distribution, on the other hand, is obtained by square-rooting Mr. Prob's credences. Mr. Bold is highly confident of contingent propositions, to the point that his credences in P and ${\sim}P$ sum to more than 1.

When we argue for Finite Additivity as a rational norm, we are arguing that Mr. Weak and Mr. Bold display a rational flaw not present in Mr. Prob. It's worth wondering in exactly what respect Mr. Weak and Mr. Bold make a rational mistake. This is especially pressing because empirical findings suggest that real humans consistently behave like Mr. Bold: they assign credences to mutually exclusive disjuncts that sum to more than their credence in the disjunction. Tversky and Koehler (1994) summarize a great deal of evidence on this front. In one particularly striking finding, subjects were asked to write down the last digit of their phone number and then estimate the percentage of American married couples with exactly that many children. The subjects with numbers ending in 0, 1, 2, and 3 each assigned their digit a value greater than 25%. If we suppose the reported values express estimates common to all of these subjects, then each of them assigns estimates summing to more than 100% before we even get to families with more than three kids!

Each of the three argument-types we consider will explain what's wrong with violating Finite Additivity in a slightly different way. And for each argument, I will ultimately have the same complaint. Finite Additivity is a linearity constraint; it requires a disjunction's credence to be a linear combination of the credences in its mutually exclusive disjuncts. In order to support Finite Additivity, each of the arguments assumes some other linearity constraint, whose normative credentials are no more clear than those of Finite Additivity. I call this the Linearity In, Linearity Out problem: We want to establish that treating credences additively is required by more fundamental rational principles. But ultimately we rely on principles that are additive (or intimately bound up with additive distributions) themselves. Instead of demonstrating that some deep aspect of rationality demands linearity, we have snuck in the linearity through our premises.

If traditional *arguments* for probabilism are question-begging in this manner, then probabilism's *applications* become all the more significant as reasons to endorse probabilistic norms. Near the end of Chapter 10 I'll ask whether Finite Additivity is really necessary for those applications. We'll briefly examine whether Bayesian epistemology's successes in confirmation and decision theory could be secured without such a strong commitment to probabilism.

Further Reading

Alan Hájek (2009a). Arguments for—or against—Probabilism? In: *Degrees of Belief*. Ed. by Franz Huber and Christoph Schmidt-Petri. Vol. 342. Synthese Library. Springer, pp. 229–51

Excellent introduction to, and assessment of, all the arguments for probabilism discussed in this part of the book.