BERT BAUMGAERTNER

DECISION THEORY

Contents

Preface and Prerequisites

Many of us are familiar with the idea of making a list of "pros" and "cons" to help us make a choice, particularly important and complex ones. Presumably, this kind of strategy reflects our desire to make the "right" choice. Or maybe there isn't a single correct choice to make, but at least we want to examine our options so that we can provide a kind of justification for what we end up deciding to do. In other words, one purpose of a pros-and-cons list is that we can provide *reasons* for our choices. Decision theory is the field that examines a whole array of decision making strategies that explain and justify choices, even if the consequences don't work out as expected.

0.1 *What is a conceptual introduction?*

This book is a conceptual introduction to decision theory. As such, there are several features that differentiate it from other introductions. First, while it draws heavily on philosophical approaches, many of the concepts are also informed by other disciplines. So when it will benefit our understanding of a concept, we will not hesitate to venture into areas such as psychology and economics.

Second, the concepts we cover will be used to build models of decision-making. In this sense the "theory" that we aspire to is not so much a thesis about what rational decision making is. Rather, our "theory" is more akin to a collection of tools that can be used for doing analysis. That is not to say, however, that we will eshew normative claims. We will find ourselves in plenty of places where we'll use claims about rationality and irrationality, and in turn the concepts we have will help us better understand those notions.

Third, there is a great deal of conceptual overlap between fields such as game theory, social choice theory, behavioral economics, and traditional decision theory (or decision theory "proper"). The differences between them is largely about the domains of decision making: game theory focuses on strategic decisions between multiple agents; social choice theory is about how to aggregate preferences of individ-

uals to the level of the collective; behavioral economics emphasizes empirical studies to develop a non-ideal decision theory; and decision theory "proper" is typically the idealized study of rationality and its normative consequences. But there are many concepts that appear across several of these domains. Rather than take a domain as a starting point and cover the concepts needed to understand the theories in those domains, we will start with concepts and then illustrate how they are used in some of these domains. That said, we will not aim to be exhaustive of where the concepts are used. The goal is for the reader to appreciate the utility of the concept and get a feel for its connections to domains outside traditional decision theory.

Fourth, to the end that we focus on concepts, we will start with bare bones. The concepts will be initially crude. At first we will use them to get some basic models of decision-making, many of which will resemble overly simplified ("toy model") decisions rather than real-life decisions. Then we will advance in one of two ways. Sometimes we will advance by adding one or more new concepts and then explore how that helps us develop more sophisticated models. Other times we will refine one or more of our existing concepts and see how that changes our models and interpretations of them.

Fifth, and finally, the goal of a conceptual introduction is to maximize accessibility. That said, it will become necessary at some point that to clarify a concept we will have to use some tools of rigor. Most readers will be already familiar with most of these: how to read a table or matrix, executing basic algebraic operations like addition and multiplication, how to read a graph that plots a function, etc. Other tools, notably some basic probability theory, readers are not expected to know. We will prefer, whenever possible, to make a concept clear without the use of formal tools. Where this is not possible, the relevant representations will be introduced as if the reader is seeing them for the first time.

The most rudimentary tools we start using early on include tables and orders of operation of addition and multiplication. Readers well-versed in these may skip to the first chapter.

0.2   How to read a table or matrix

This books make very few assumptions about the reader's background knowledge. This book makes frequent use of tables, or sometimes called matrices ("matrix" for a single one). So it's worth getting clear on the parts that make up a table. The basic setup is a grid, with horizontal lines (left-right) and vertical lines (up-down). Sometimes the lines aren't shown because it looks clearer. In that case we make sure to organize text in a way that lines up as a grid.

The horizontal regions are called rows and the vertical regions are columns. A cell is where a row and a column intersect. The cells in the very first row usually contains information about the names of the columns (this first row is often called a *header*). The cells in the very first column typically have the names for the rows. Because the first row and first column are being used to label or name the rest of the row or column, they aren't included when we count the number of rows (or columns) of a table. Below is an example. Notice that the whole grid is made up of three rows and three columns, but because the first ones just have names, the counting starts at the second row and column of the grid.

Table 1: Example of a two-by-two table with names for rows and columns. The name "Row 1" refers to the horizontal region to its right and consists of Cell(1,1) and Cell(1,2). The name "Column 1" refers to the vertical region below it and consists of Cell(1,1) and Cell(2,1).

	Column 1	Column 2
Row 1	Cell(1,1)	Cell(1,2)
Row 2	Cell(2,1)	Cell(2,2)

0.3   The Very Basic Math

The math that we will use is itself not advanced. It basically amounts to the operations of addition, subtraction, and multiplication (and in a few exceptions division). But what can become a bit tricky is the "book-keeping" of it all. This book will generally show all the work so that you can follow the order of operations.[1]

[1] If you need a refresher, do some practice problems.

0.4   Inspiration and Acknowledgments

This text is inspired by two freely available sources.[2] One is the open access book *Odds & Ends: Introducing Probability & Decision with a Visual Emphasis* by Johathan Weisberg `https://jonathanweisberg.org/vip/`. The second is Brian Weatherson's Decision Theory Book `http://brian.weatherson.org/424/DTBook.pdf`. There is a great deal that I like about both of these, but they did not quite fit the needs I had when teaching an introductory course on decision theory. While both are relatively approachable, I wanted to emphasize some of the concepts that I thought students are most likely to remember. Most of my students shy away from technical material. My goal was to assume as little as possible and very gradually introduce them to

[2] Students, you can skip the rest of this and move to the first chapter.

the formal methods needed, and only when needed.[3] I also wanted to emphasize related topics, such as a broader conversation about the nature of utility and the perspectives of behavioral economics, both of which are understandably not given much attention in more standard philosophy texts.

On the technical side of things, this book was made using the bookdown package created by Yihui Xie [@xie2015] and inspired by the formatting aesthetics of Tufte.

1

Introduction

OUR ACTIONS are the consequences of our beliefs blending with our desires. I buy some ice cream because I have a *desire* for something sweet and *believe* that ice cream is a good means of satisfying that desire. I take the train to shop downtown because I *believe* the train is an efficient means of transporting me there and I *want* to purchase some things. This way of talking about our actions and choices is sometimes called "belief-desire psychology" and it is a helpful big-picture way of dividing up the parts that make up a decision: we consider a set of possible actions and do the action that best combines our beliefs and desires. If I don't believe the train will get me downtown, or if I don't want to go shopping after all, I won't take it.[1] This picture is misleadingly simple. This chapter gives some brief introductory concepts to refine our understanding of decisions so that we can, over the course of this text, improve our analyses of them, and hopefully improve them.

[1] In this way our beliefs are like bets on what the world is (going to be) like. Frank Ramsey (1903-1930), a founder of modern decision theory, suggested this kind of picture.

1.1   *Some Basic Conceptual Ingredients*

Let's start with an example. Suppose you're thinking about how to spend your Sunday afternoon. You've got a project due on Monday, a project that you get at least some enjoyment out of working on. If you had to, you know you can get the project done quickly, but the time crunch would suck the enjoyment out of doing it. You also know that if you go to the park there will be a pick-up game of frisbee. Your favorite pick-up games have a mix of friends and new people, but you really don't like it when that annoying guy shows up. If you went to the park and a game never came about, or that annoying guy showed up, you'd regret not having spent your afternoon working on the project.

A DECISION TABLE (or decision matrix) can help us organize the

above example.[2] We use rows to lay out your choices, the options you are considering. In this case the choices you're thinking about are going to the park or working on your project. We use columns to represent features that are outside of your control, what we call states of the world. Using the above example, let's say the possible states of the world are a fun game of frisbee and a game of frisbee with annoying guy. The cells of the matrix are called outcomes and we can insert numbers to indicate how much you like each outcome. Let's assume that higher numbers mean that you like that outcome more. Here's one example of a decision matrix (we'll sometimes call it a "decision table").

	Fun Game	Annoying Guy
Go to park	4	1
Work on project	2	3

[2] A decision table uses rows for possible actions and columns for possible world states.

Here's how to interpret this basic 2x2 decision table.[3] There are two options represented in the rows and two world states represented in the columns. Each number in a cell represents how much that outcome is liked relative to the other outcomes.[4] Here, the outcome in which you choose to go the park and the frisbee game turns out to be fun has the highest number (4) and so is the most preferred outcome. The least liked outcome is the one where you choose to go to the park and annoying guy shows up and ruins the game (1). The outcome where you chose to work on the project and you find out later that annoying guy did not show up and your friends had a fun game is better than the worst outcome (because 2 is higher than 1), but it's worse than the outcome where you worked and annoying guy showed up (because 2 is lower than 3).

Why do we have this order of preferences for the outcomes? There could be a variety of reasons. Maybe if you're working at home on your project, you'd rather find out that annoying guy was there. After all he's a pretty good player and your friends don't mind him as much as you. Or perhaps your friends find you more annoying than annoying guy and you recognize this. Whatever the reasons, the decision table does not represent those details. The table just represents the outputs in terms of ordering the preferences, using higher numbers to indicate more preferred options.

Preferences can differ. For example, you might think it's better that both you and annoying guy didn't show up than the outcome where only you didn't. (Perhaps you are less annoying than annoying guy after all!) To represent this, we would change the numbers in the bottom row of the matrix:

[3] Not every table is a decision table, so pay careful attention to *what* a table is intended to represent.

[4] In general, the number in a cell by itself doesn't tell us anything; it is in comparison with other cell numbers that we get information that decision theory makes use of.

	Fun Game	Annoying Guy
Go to park	4	1
Work on project	3	2

Note that both matrices represent just how much you dislike annoying guy: your most preferred outcome is where you go to the park and he's not there, but if he does show up you're now in your least preferred outcome. In fact, in both matrices you'd rather be working on your project then go to the park where he shows up. But in his absence you have a fear of missing out and you'd rather go to the park than work on the project.

This example illustrates the first set of concepts we will use to start modeling decisions. In the simplest decision we need to have at least two choices and two states. Sometimes we'll call a choice an option or an action that you can pick. We'll use these terms interchangeably until we need to be more precise. The variables a, b, c, and d are the numbers used to measure how good the outcomes are. We will have much to say about these in what's to come.

	State 1	State 2
Choice 1	Outcome a	Outcome b
Choice 2	Outcome c	Outcome d

Real life decision-making often has more than two choices. To represent that we would add more rows, one for each choice. Similarly, there can be more than two world states, and so we would add more columns. The number of choices need not match the number of world states.

Sometimes the states of the world will depend on the actions of others, sometimes not. In our example, the state where annoying guy shows up to the park is itself an outcome of a decision that annoying guy made. For all you know, he might find you really annoying and is making similar considerations about whether he should stay home and work on his project. Ideally you could coordinate, so that you alternate from one week to the next so you don't have to see each other. Decisions where the outcomes depend in large part on the choices that others make fall under the field of *Game Theory*. A classic example of this would be Rock, Paper, Scissors. Let's say it's 0 for a tie, -1 when you lose, and 1 when you win. We can use the matrix representation to organize the options, states, and outcomes. Suppose you are playing with your friend Joe.

	Joe plays Rock	Joe plays Paper	Joe plays Scissors
You play Rock	0	-1	1
You play Paper	1	0	-1
You play Scissors	-1	1	0

Sometimes decisions are made by aggregating choices or preferences. For example, your friends may take a vote for who is more annoying, you or annoying guy. Elections are another example of how a group decides who the next political official will be. The study of how to aggregate lower-level decisions into higher-level decisions is known as *Social Choice Theory*.

Many of our decisions don't depend on the strategies of others, nor are aggregates. Many decisions either don't involve other people at all, or the actions of others are largely independent of yours (and vice versa). For example, your going to the park to play frisbee may be primarily concerned about whether it will rain or not. Whether you decide to learn to play an instrument may depend primarily on whether you think you will enjoy it. The decision to have children may largely consider the type of experience you can expect to have in a parental relationship. Choosing to go in for surgery will depend a lot on the likelihood of success. Buying some new hand cream will depend on whether you think the cost is worth the amount it improves your dry skin.

As you might guess, not only can decisions be complex, but the study of decision making is itself complicated. In fact, decision theory broadly construed is an interdisciplinary field, involving researchers from areas such as philosophy, psychology, political science, economics, computer science, and business. It is a rich and diverse field, but we have limited time and space. So we're going to approach the topic of decision-making from a mostly philosophical perspective that emphasizes the concepts used to model decisions. But we won't shy away from making connections to other areas or from dealing with messy topics. Perhaps the messiest of all is about *rationality*, which we'll briefly cover next.

1.2   Rationality - the Descriptive and Normative

A DESCRIPTIVE THEORY is an attempt to characterize how decisions are made in the real world. It draws heavily on sciences that are traditionally empirical, including psychology and economics. There's a entire subfield devoted to it called behavioral economics. And perhaps surprisingly, there's a subfield in philosophy too that uses

empirical methods under the rubric of experimental philosophy. Descriptive decision theory is interested in explaining how humans make decisions.

A NORMATIVE THEORY aims to provide prescriptions of what people *ought* to do as rational agents.[5] Failing to meet the prescriptions of a normative decision theory means running the risk of being labeled irrational. Less harshly, if we fail to meet the conditions of rationality, we shouldn't be surprised if our choices and lives don't go as hoped. Normative decision theory is interested in decision making that is justifiable.

[5] Sometimes "prescriptive" is used instead of "normative" - we'll use these terms interchangeably.

Both descriptive and normative decision theory provide important contributions to our understanding of decision making. The normative side allows us to make recommendations of how to make decisions and can provide guidelines for attributing responsibility when someone fails to meet the ideal - we might for example want to be able to say when someone is making bets that they shouldn't be. On the other hand, we often will want to provide an explanation of how someone came to a decision, even if it failed to meet the ideal. Why did Jodie think (mistakenly) it was a good idea to keep betting on the horse that kept losing? Here we are asking for an explanation of Jodie's thinking, not for a justification of it.

While it may seem that we can carve decision theory into these two sub areas, the truth is that they very much inform one another. A commonly held principle is that "ought implies can''. For example, if it is true that Sally ought to do her homework, then it is also true that it is possible for Sally to do her homework. We can actually restate this principle in a logically equivalent way (what logicians call the contrapositive) as follows: if it is impossible for Sally to do her homework (perhaps she is in a medically induced coma) then it is no longer true that she ought to do it.[6] So prescriptive demands on our decision making have to respect what is possible for us to do.

[6] In general, the *ought implies can* principle says that normative claims cannot "outrun" what is possible.

In brief, descriptive decision theory provides us with an understanding of what we do and what we can do with respect to making decisions. On the other hand, what we do or can do is not necessarily justified and the role of normative decision theory is to set up one or more ideals by which we can compare what we do in order to become better. There's much more to be said about the interaction between the normative and descriptive, but we'll leave it at that for now.

In both descriptive and normative decision theory, the notion of rationality that is being studied is *instrumental* rationality.[7] Here's what that means. It is presupposed that agents have some aim or goal. Decision theory is not about what aims we have - goals are so

[7] "Instrumental" that is, in the sense of means to some end, where the "end" can vary across persons, groups, etc.

to speak "external'' of the theory. Rather, decision theory is about characterizing the process for how to best ensure that one's goals are satisfied. Some philosophers argue that rationality can go beyond just what is instrumental, that is to say, that we can argue about which goals are rational to have and which are not. John Rawls, for example, argues that counting blades of grass on a courthouse lawn is simply not important enough to count as a rational aim. Other philosophers argue that all rationality is instrumental, that the only way to evaluate a goal is by looking at how well it does to satisfy some other overarching goal. There is no way, these philosophers argue, to scrutinize the most fundamental or overarching goals. Most of decision theory sidesteps this debate. It proceeds by asking how best to proceed *given* an aim, whatever that aim might be and whether that aim is rational or not.

To study instrumental rationality is to study the process by which we have the best chance to obtain our aims. It does not, however, guarantee that those aims will be achieved. Recall that outcomes are a combination of actions and things external to the decision that are outside our direct control (states of the world). Consequently, this means that the best decision could turn out to be the wrong one retrospectively, and a second rate decision could turn out to be the right one. These possibilities are by-products of the fact that decisions often have to be made by a point in time. For example, while camping John may drink water that came from a filter that clears Giardia. To the best of his knowledge the filter works and the alternative to drinking the water is to get severely dehydrated. So it is instrumentally rational that John drink the water. However, unbeknownst to him, the filter has a manufacture defect. So the resulting beaver fever suggests that his decision was not the right one - dehydration, while uncomfortable, is easier to recover from the malladies of Giardia. Note that the actual outcome does not change the fact that John made the best decision given that the alternative was severe dehydration. This is because decision theory operates on the basis that decisions are made with information available by a certain time and not information made available at a later point in time. Notice here the *ought implies can* principle is at work: John can't be faulted for not knowing the future. This brings us to the topic of uncertainty.

1.3   *Uncertainty*

In some cases we know what the state of the world is or will be. When we make decisions based on knowledge, it is called *decisions under certainty*. For example, suppose you have $20 to spend on lunch and you want pizza. You have stale leftover pizza at home, but it's a

little further away than Maialina, where the pizza is fresh out of the oven. On the other hand, you don't mind stale pizza. Most importantly, your preference for not spending $15 is much stronger than eating fresh pizza and having to walk a little less further, so much so that the other considerations are irrelevant to you. A first draft of a decision matrix to model this decision might look like this:

	World State
Go Home	$20 in pocket and stale pizza
Go to Mailalina	$5 in pocket and fresh hot pizza

If, as we supposed, the differeence between stale and fresh pizza is negligible and your primary concern for your decision about what to eat is money, then all that matters for this particular decision is your preference regarding the outcomes. This helps to simplify our model of the decision, because we just need a number for each outcome and make sure that we give the higher number to the outcome we prefer:

	World State
Go Home	2
Go to Mailalina	1

Clearly, you should go home and eat the leftover pizza, since 2 is higher than 1.

Not all decisions under certainty are this simple. Suppose you want to redecorate your living room. You aren't going to buy anything new, but you might consider removing some things (you're not attached to that painting you did in 9th grade anymore). What you're thinking about is how to *optimize* the location of the various pieces you have. *Optmization problems* are typically examples of decision under certainty. Construction companies do this. Even when they know that the price of gravel from GoodGuysGravel is cheaper than GreatGuysGravel, they have to combine that with knowledge about their respective delivery charges. GreatGuysGravel delivery charges may decrease the more gravel you order, so that higher quantities of gravel turn out to be less expensive to get than the competitors, who charge the same amount for delivery no matter how much gravel you want. In brief, this is an example of decision making under certainty because all the relevant information is known and the remaining task is to minimize cost.

A lot of decision making happens when we have less knowledge of the world. When we have some information about what will happen but it's not perfect, this is called *decisions under risk*. An impor-

tant feature is that we can assign probabilities to outcomes. The best examples are in gambling where the possible states can be clearly and carefully quantified. The probability of rolling two ones (Snake Eyes) with a pair of die is 1/36 or 2.77%. We know this because we can count all the combinations of rolling two die (there are 36) and only one of them is a pair of ones. The probability of rolling a seven is 16.67% because of the 36 possible combinations, in six of them the two numbers add up to seven. Sometimes these kinds of situations are called *known risk* since the outcomes are chancy, but we know what those chances are.

At another extreme, we might have no information that helps us assign probabilities to outcomes. These are called *decisions under igno-rance*. I'll sometimes refer to such situations as having *unknown risk*. For most of us, the exact date of our death is like this, and many of us would prefer to keep it that way. Another example could be the outcome of telling a joke at a party with strangers. It might generate big laughs, fall flat, or may even be offensive. Without information about who these people are, you don't know how to assign probabilities to the outcomes.

Lots of decisions are in between the extremes of known risk and unknown risk (ignorance). That is, we have some information so that we're not completely ignorant, but we don't have so much informa-tion that we can carefully and explicitly assign probabilities. Some presentations make a clear delineation between risk and ignorance, and will sometimes call the latter cases of uncertainty. Here we're going to use *decisions under uncertainty* as an umbrella for the whole continuum that goes from risk to ignorance.

The Allegory of Fortune by Dosso Dossi (1486-1542). Fortuna ("Lady Luck") can bring good luck (fruit), but good luck tends not to last (she's sitting on a bubble). She might also bring bad luck (missing sandal). Chance (on the left) is holding lottery tickets. Art critics say Dosso Dossi believed life is a lottery for everyone.

1.4   Practical and Theoretical Problems

One of the riches of studying decision-making is that the same con-ceptual tools can be applied to both practical and theoretical prob-lems. Practical problems range from ones we experience in everyday life to the kinds we see on game shows. Theoretical problems are ones that no human is expected to actually face, but by thinking about how someone in such a theoretical situation would solve the problem we can better understand how to approach practical situa-tions.

A FAMOUS PRACTICAL DECISION PROBLEM is *The Monty Hall Prob-lem* which is named after the host of the game show *Let's Make a Deal*. On the show there are three doors, one of which with a prize behind it. You get to pick one of the doors. Let's say you pick A. The host now opens one of the other two door that you did not pick. But of

Monty Hall was creator and host of the 1960's game show *Let's Make a Deal*, which was rebooted with Wayne Brady (not pictured).

course, the host doesn't want to give away the game, so the door they open will be empty. Let's say they opened B and sure enough it's empty. Now the host asks, do you want to switch your choice to door C or stick with your current choice of A? Take a moment to think what you would do.

Most people have the intuition that switching your choice makes no difference. That is, that both doors A and C have equal chances of having the prize: each is 50%. That is even what many professors of statistics and mathematics find sensible. But the right answer is that you switch! After the host opens door B, you are twice as likely to win the prize if you switch your choice to C than you are sticking with your choice. The reasoning is subtle and we'll need to use concepts from probability theory, which we will get to soon enough. But to give you a feel, ask yourself: if instead of three doors there were 100 doors, and the host closed 98 of the doors, would you switch your choice then? Or imagine two people playing the game many times, one person that always stays with their choice, the other always switches. Who is expected to win more often in the long run?[8]

A FAMOUS THEORETICAL PROBLEM is *Newcomb's Problem*. Suppose you are on a game show and there are two boxes in front of you: A and B. The contents of box A are concealed, but box B is completely transparent. Inside of transparent box B is $1000. The options you are given might seem a bit strange, but this is what they are: take the unknown contents of box A (this is called the 1 box option), or to take the contents of both box A and box B (this is the 2 box option). Here's the twist, and this is what makes this a theoretical problem. While the contents of box A are unknown to you, there is an extremely sophisticated artificial intelligence that is very good at predictions. We'll suppose this AI has made perfect predictions every time in the past. The game show host consults the AI and does the following. If the AI predicts that you will pick the 1 box option (i.e., you pick the unknown contents of A) then the host will put in $1,000,000 in box A. If, on the other hand, the AI predicts that you will pick the 2 box option, then the host will leave box A empty. All of this was already done before you are presented with the options. We can use a decision matrix to represent the example:

	AI predicts two box	AI predicts one box
One box (just A)	$0	$1,000,000
Two box (A and B)	$1,000	$1,001,000

Here's one way of reasoning about what to do. The fact is that

there's either a million dollars in box A or there's not, your choice won't make the difference. If there is a million dollars in box A, then picking the two box option is better than the one box option. After all, $1,001,000 is more than $1,000,000. If box A is empty, then the two box option means that you'll get at least $1,000, which is better than $0. Either way, the two box option is better.

As compelling as that argument might be, here's another line of reasoning that gets us to the exact opposite conclusion. You know that the AI has an incredible track record. Of the hundreds of people on the game show before you, the AI has gotten all of the predictions correct: those who picked the one box option got one million dollars, and those who picked the two box option got one thousand dollars. You want to be like those in the first group and get a million dollars rather than those in the second. So, you should pick the one box option.

Which of these arguments is better?

Earlier I claimed that thinking about theoretical problems can help us approach practical situations. So, are there any applications of Newcomb's problem?

A central feature of Newcomb's problem is that there is something in the world that is tracking decision making processes that then leads back into a decision problem. In the fantastical case it was an AI making predictions. We can remove this fantastical part and think instead of examples where there is a common cause between the world being in a certain state and you making a choice. Let's look at a version of the example we started this chapter with.

One of the things that makes annoying guy so annoying is that he loves to talk about the projects he's working on, projects you just don't care about. Come to think of it, you realize that you are actually quite similar to annoying guy: You also like working on projects and talking about them. That's why your friends find both of you annoying, especially if both of you show up. They don't mind having just one of you there, though (they have just enough tolerance for one annoying person, say). We can set up a decision matrix that looks like this:

	Annoying guy stays home	Annoying guy shows up
You stay home	2	0
You show up	3	1

If you knew that annoying guy stayed home, you'd prefer to show up (3 is better than 2). If annoying guy showed up, you'd still rather be playing in the park than be at home (1 is better than 0). That said, if you both stayed at home, that would be better than both showing

up (2 is better than 1). So what should you do?

Here's one line of reasoning that resembles the argument for the two box option above. Either annoying guy is going to stay home or show up. If he stays home, you prefer to show up and play. If annoying guy shows up, you still prefer to show up. So whether annoying guy stays home or shows up, you prefer to show up. So you should decide to go.

What the above argument does not consider is that annoying guy is just like you. That means whatever you decide to do will be a good prediction of what annoying guy decides to do. In fact, we could suppose that his decision matrix is identical to yours, except that in his mind you are the annoying one. It may help to explicitly represent things from annoying guy's perspective by transposing the above table as follows.

Table 1.9: From the perspective of annoying guy, his table would look like this.

e	You_stay_home	You_show_up
He stays home	2	3
He shows up	0	1

So here's an alternative line of argument. Because of how similar you two are, there's little chance that one of you will show up while the other doesn't. In other words, it's much more likely that you both choose the same thing and you both show up or you both stay home. You prefer the outcome of both staying home than both showing up (2 is better than 1). So, you should decide to stay home, and that's likely what annoying guy will choose to do as well (since he'll be reasoning similarly). Your friends couldn't be more pleased.

The key difference between these arguments is whether you are using information about the decision you are making in assessing the probabilities of the outcomes, or whether the probabilities of the outcomes are independent of your choice.

Later on we will see how these arguments can be made more precise. There will be two ways of specifying the probabilities of outcomes, one that does not conditionalize on choices, and one that does. In addition, we will see how the fact that this crossroad exists at all means that the two most popular principles of decision making can come apart even though they agree most of the time. One is called *the dominance principle* and the other is called *maximizing expected utility*. But let's not get too ahead of ourselves. First we'll turn to some decision making concepts that rely on what we have implicitly been assuming: that preferences form ordinal rankings.

1.5   Summary

The most basic model of a decision has the following four ingredients:

1. At least two exclusive choices/options/actions, which we represent as rows in a table.
2. At least two states of the world (that are out of our control), which we represent as columns.
3. Outcomes, the cells in the table.
4. Preferences, the numbers we assign to the outcomes, where higher numbers represent outcomes we prefer relative to others.

Rationality can be understood *descriptively* and *normatively*.

Some decisions are made under certainty, but most of the ones we'll study are under uncertainty.

Decision theory, as we'll study it, draws from both theoretical and practical considerations.

Exercises

1. If there are two possible world states and three possible actions, how many possible outcomes are there?

2. What are the rows in a decision table?

    a. Things that you can control in the decision
    b. Things that are outside of your control in the decision
    c. Things you wish you could contorl but you can't.
    d. The outcomes of your decision.

3. Suppose Dr. Smith is conducting an experiment on whether people prefer one marshmallow today or two marshmallows tomorrow. Which of the following best describes the field that Dr. Smith's experiment is contributing to?

    a. Descriptive decision theory, because Dr. Smith is getting information about how many people choose the correct option: one marshmallow today.
    b. Normative decision theory, because Dr. Smith is using normative claims and testing which ones are true.
    c. Descriptive decision theory, because Dr. Smith is collecting information about how people make a choice between present and future outcomes.
    d. Normative decision theory, because there is no correct answer.

4. Normative decision theory aims to describe. . .

   a. how things are.
   b. how things ought to be.
   c. how decisions ought to be made.
   d. how decisions are made.

5. You're at the Kentucky Derby deciding on whether to bet on this
    year's favorite horse: The Decider. The probability of The Decider
    winning is 1/7. Which of the following describes the kind of deci-
    sion you're considering and why?

   a. Decision under known risk, since there is a body of evidence,
       like past races, that supports the claim that the probability that
       The Decider will win is 1/7.
   b. Decision under unknown risk, since nobody can be certain that
       The Decider will win.
   c. Decision under certainty, since we know that the chances of The
       Decider winning is 1/7.
   d. Decision under unknown risk, since there is a larger body of
       good evidence that the other horses haven't done very well
       when racing against The Decider.

6. In various forms of gambling like roulette or poker, one can assign
    probabilities to the outcomes and make a prediction accordingly.
    This is known as a. . .

   a. Decision under risk
   b. Decision under uncertainty
   c. Decision under ignorance
   d. Decision under the influence

7. You're at a small funeral with people you know very well. You,
    like everyone else, knows for certain that death is coming for us
    all, but we are uncertain about when that day will be. You're de-
    ciding between three options: i) say some kind words about the
    deceased, ii) say some honest but unkind words about the de-
    ceased, and iii) say nothing at all. What kind of decision scenario
    is this most like?

   a. Decision under deceased influence.
   b. Decision under certainty.
   c. Decision under ignorance.
   d. Decision under risk.

8. Suppose you are on a game show where they have the same setup
    as Newcomb's problem. The host has already consulted with the
    devil and chosen to either put the money in Box A or not. Just

as you are about to make your choice between the one box and two box option, you happen to get a trustworthy glimpse of the contents of Box A, which turns out to be empty. What option do you pick?

  a. Two-box option
  b. One-box option

9. Suppose you are on a game show where they have the same setup as Newcomb's problem. The host has already consulted with the devil and chosen to either put the money in Box A or not. Just as you are about to make your choice between the one box and two box option, you happen to get a trustworthy glimpse of the contents of Box A, which turns out to **have $1 million**. What option do you pick?

  a. Two-box option
  b. One-box option

10. True or false: in the Monty Hall problem, it's essential to the puzzle that the host doesn't want to expose the prize. If they didn't care about giving away the location of the prize, there would be no reason to switch when they open door C.[9]

[9] This question comes from *Chapter 1* of Jonathan Weisberg's *Odds and Ends*.

11. A characteristic feature of game theory is what?

  a. The rows are options that both you and someone else jointly control.
  b. The world states are the actions of someone else.
  c. It is decision theory applied in the domain intellectual games, including the solving of a Rubik's cube.
  d. Game theory is not a real thing. You made this up.

2
Ranking

This chapter will be on decision making that relies primarily on rank-
ing outcomes. So far we have been using numbers to represent how
much we like an outcome, where higher numbers represent outcomes
we prefer more than outcomes with lower numbers. More specifi-
cally, we have been using what are called *ordinal utilities*[1] where all
we care about is whether one outcome is better, the same, or worse
than another outcome. That is, we aren't paying attention to how
much more we prefer one outcome to another. We'll get to that later.
For now, if there are four outcomes and we give the one we prefer
the most a utility of 4, we could have also given it a utility of 400. All
that matters for ordinal utilities is that they rank outcomes from best
to worse (with possible ties).

> [1] **Ordinal utilities** make qualitative comparisons (x is better/same than y) but not quantitative comparisons.

    To get a feel for what decision making rules look like, we'll start
with some simple ones. Because of their simplicity, they may not
seem like very plausible rules and so we won't ultimately suggest
that these should be followed. But understanding what the simpler
rules get wrong is informative and will help us get a better idea for
who to develop more sophisticated decision making rules.

2.1   Maximin and Maximax

Suppose you're a student that wants to have it all: you like going
to parties, but you also take your studies seriously. Tomorrow there
will be an exam, but you don't know if the professor has designed
a difficult exam or an easy one. If you knew that tomorrow's exam
were going to be an easy one, you'd prefer to go party, but if you
knew that the exam were going to be a difficult one, you'd prefer to
study and be prepared for it. And "all things being equal" between
partying and studying, you have a preference to study because you
can't stand the thought of getting a C grade - which is a possibility
if the exam is a difficult one and you didn't study. Your preference
assignment (pun intended) might look like this:

	Difficult Exam	Easy Exam
Party!	1	3
Study!	4	2

What should you do?

Here's one decision rule you might use: Maximax. Maximax is a rule that says you should **max**imize the **max**imum outcome. The maximum utility in the above table is 4 and is assigned to the outcome in which you study and the exams turns out to be difficult. Of course, you can't control whether the exam will be a difficult one or an easy one (that's up to the professor), but Maximax doesn't really care. It just says you should pick the action that has the outcome with the highest utility.

Maximax isn't a very plausible rule if you care at all about risk. It has you choose outcomes with the highest payoffs, but with complete disregard for any possible outcomes with lower utilities. At a horse race Maximax says to bet on the drunk horse that's missing a leg because the payoff would be so much bigger than any of the other horses that clearly will leave the drunk one in the dust. Yes, it's possible that all the other horses suddenly have heart attacks at the start, but that possibility is so remote that you would be right to ignore it. At the very least, we should be incorporating information about the chances that the drunken horse will win, which are very, very low. Maximax doesn't do that. All that matters to Maximax is the size of the payoff. Nothing else.

Maximin is a much better rule. Maximin says to choose the least bad worst case scenario, i.e., **max**imize the **min**imum outcome. What you do is find the worst case scenario for each option, and then select the option with the least worst case. In the table above, the worst case scenario for the Party! option is 1 and for Study it is 2. Since 2 is greater than 1, Maximin says to study.

So far Maximax and Maximin agree, but they won't always. For example, suppose that your preference assignment is captured by the following table:

	Difficult Exam	Easy Exam
Party!	1	4
Study!	3	2

Now Maximax and Maximin disagree. Maximax says to party, since it has the outcome with the highest utility (4). Maximin, on the other hand, says to study. Why? The worst outcome if you choose to party is 1, while the worst outcome if you choose to study is 2.

Between those two, 2 is better than 1, and that can only happen in the choice to study.

Maximin is a pretty good rule if you are risk averse, that is, if you are predominantly worried about decreasing negative consequences.[2] Not everyone may be as concerned about risk aversion as others. There are at least two ways of thinking about this. One is to understand differences between people's decision making as differences in the rules or strategies they use to make a choice. Another way to understand differences between people is to recognize differences in people's preferences. For example, if you're of the mindset that "C's get degrees" and you really love to party, then you might set up your preferences like this:

	Difficult Exam	Easy Exam
Party!	3	4
Study!	1	2

For someone with this preference assignment, both Maximax and Maximin recommend that they should party. So the difference between why someone studies and why someone else parties might be a difference in preferences, or it might be a difference in the decision rules they're using. We'll turn to this point more shortly.

Maximin isn't without its own limitations. It can't, for example, help us decide what to do in certain situations where there are ties between outcomes across different options. Suppose someone is indifferent between all outcomes except the one where they studied and the exam was difficult. They might have the following preference assignment:

	Difficult Exam	Easy Exam
Party!	1	1
Study!	2	1

Maximin doesn't make a recommendation about what to do because there is a tie between the worst outcomes of the two options: the worst in each case is 1. And yet, you might have the intuition that someone with this preference assignment should choose to study. But the reason, you might think, is not because they should use the Maximin rule. Rather, it's something about the relative comparison between partying and studying across the different possible world states. If you have this intuition, you'll already have a feel for the next decision making rule we'll explore: *dominance reasoning*.

2.2   *The Dominance Principle*

The dominance principle does a more sophisticated comparison between options. There are two versions, a *weak* version and a *strong* version.

Suppose there are two options, A and B. We say that option A *weakly dominates* option B if, for every world state, A is at least as good in its outcome as B. In other words, what we do is this. We check every column to make sure that A is at least as good as B. If so, then A weakly dominates B. If not, then A does not weakly dominate B. Let's look at same table again:

	Difficult Exam	Easy Exam
Party!	1	1
Study!	2	1

Under "Difficult Exam" the option to study is at least as good as partying (2 is better than 1) and under "Easy Exam" the option to study is also at least as good as partying (1 is the same as 1). So, by the definition of weak dominance we just gave, the choice of studying weakly dominates the choice to party. Notice that we are *not* doing a comparison between options across two different states (e.g. bottom left and top right), our comparisons are between options in the same column.

The Weak Dominance Principle says to *never pick weakly dominated options*. In the above table, studying weakly dominates partying, or in other words, partying is weakly dominated by studying. So the Weak Dominance Principle recommends that you do *not* party. Since there are only two choices in our example, by the process of elimination the only option you are left with is to study.

We say that option A *strongly dominates* option B if in every state option A is *strictly better* in its outcome than B. Notice that in the above table, while it is true that under Difficult Exam studying is strictly better than partying (2 is higher than 1), the same is not true under Easy Exam - here both options are equal. That means that studying does not strongly dominate partying. So, it is possible that an option can weakly dominate another, but not strongly dominate it.

But let's suppose that the preference assignment was just a little bit different for the outcomes under Easy Exam:

	Difficult Exam	Easy Exam
Party!	1	3

	Difficult Exam	Easy Exam
Study!	2	4

Here the option to study both weakly dominates the option to party, and it strongly dominates it. In fact, whenever an option is strongly dominated by another option, it is also weakly dominated by it.

The Strong Dominance Principle says to *never pick strongly dominated options*. In the table just above partying is strongly dominated by studying (which means partying is also weakly dominated by it), so the Strong Dominance Principle says to not pick the option to party. Here both the strong and weak versions of the Dominance Principle agree.

Notice again that it doesn't matter that the best outcome for partying (3) is better than the worst outcome for studying (2). The two options are being compared column by column (i.e. state by state). Notice also that the Dominance Principle, or "dominance reasoning", agrees with the recommendations of Maximin and even Maximax.

2.3   *More than two options and two states*

The decision making rules we've explored so far can be generalized to situations where there are more than two options. It can be helpful here to think a bit more abstractly. Let's say there are three options, A, B, and C. And suppose there are two world states, 1 and 2. Then we need a table with three rows and two columns. Finally, let's fill in the outcomes with the following utilities:

	World State 1	World State 2
Option A	2	4
Option B	1	3
Option C	5	6

Here it becomes apparent why we stated the Dominance Principle the way we did. We said to *never pick dominated options*. Notice we didn't say "pick an option that dominates another option". This is because it is possible for an option to dominate another, but be itself dominated by some third option. By saying "never pick dominated options" we ensure that the option is not bested by some other one. So while in the example above it's true that A dominates B, it is also true that A is dominated by C (i.e., C dominates A). Option C, however, is not dominated by any other option. In fact, it's the

only option that isn't dominated by another, and hence that is what dominance reasoning recommends.[3]

Notice that in this example all of our decision rules so far agree: they all recommend Option C. But the *reason* for their recommendations are all different. For Maximax it is because C has the outcome with the highest utility of 6. For Maximin it is because C has the least worst outcome with a utility of 5. And for Dominance Reasoning it is because it is the only option that is not dominated by any other. We will leave it as an exercise to ponder whether these reasons can ever make different recommendations.

You may also notice that none of the rules we've used are limited to decisions that have only two world states. This is most obvious in the case for Maximax: just look for the outcome with highest utility, follow that row to the left, and then pick the option that corresponds to that row. For Maximin you look for the lowest utility in each row, and then pick the option that has the highest one of these. Dominance reasoning is a bit more complicated, but not by much. When there are two columns, you are comparing two options to see if one is dominated by another, and you do this by checking the first column and then the second. In a decision matrix that has three columns you do the same thing, but you have to compare the two options in the additional third column. In other words, each additional world state means that you have to check an additional column when you're asking whether one option dominates another.

2.4   Non-Unique Recommendations

We mentioned earlier that there are at least two different explanations for why two different people might make different decisions, even if they have the same options and the same world states in their decision tables. One explanation is that they could be using different decision rules: maybe one person is using Maximax while another is using the Dominance Principle. We know that these rules can make different recommendations, so perhaps that explains the difference in two people's choices. (Question for reflection: can someone using the Dominance Principle come to a different recommendation than Maximin?)

Another possible explanation for differences in decisions is a deference in preferences. We've seen this already as we considered different preference assignments. For example, one student might have the mindset that "C's get degrees" and very much enjoy a good party. So they might have their preferences like this:

[3] You should make sure you can do this reasoning on your own. It's helpful to focus on just two options at a time, and then striking one out if it is dominated.

	Difficult Exam	Easy Exam
Party!	3	4
Study!	1	2

A student with this above preference assignment will come to the conclusion that they should party, regardless whether they follow Maximax, Maximin, or the Dominance Principle.

Another student might have post-graduation ambitions that require higher grades than C. Moreover, they might not be fond of parties much anyway. So they might have a preference assignment like this:

	Difficult Exam	Easy Exam
Party!	1	2
Study!	3	4

The student with this above preference assignment will come to the conclusion that they should study, regardless of whether they follow Maximax, Maximin, or the Dominance Principle.

In other words, the answer of what to do is not unique, even if two people agree on a decision rule. Different people could have different preferences, which can lead to different recommendations. So rationality, in so far as we have been considering dominance reasoning, does not demand that we all make the same choice. But be careful, that does not mean that we can choose anything we like without being irrational - we have to be mindful of what our aims or goals are. For example, while it may be rational for the first student to go party (after all, that's what dominance reasoning tells them to do), it would not be rational for the second student to go party. Our intuitive judgment suggests this is because going to the party is at odds with the goals of the more ambitious student. The dominance reasoning strategy gives us the same result given that the preference assignment in the second decision matrix reflects the aims of that student.[4]

[4] This is what we had in mind by *instrumental* rationality.

Notice that dominance reasoning proceeds by comparing options. We will see with other rules that there are alternative ways of making comparisons. But at the heart of most rules is the idea that options are assessed by how they are expected to perform relative to each other. We will see scenarios where this feature of decision making can get us into trouble and will try to find fixes for that.

It's important that you don't confuse the comparative nature of decision making with relativism. Relativism is a position about the nature of truth, particularly truths regarding ethical claims. While there are some connections between the normative aspects of decision

theory and the study of morality, decision theory makes no commitments about the nature of truth. Well, almost none. We'll get to that later when we consider the nature of probability. But for now, make sure that you recognize that while we are working with the understanding of rationality in the instrumental sense (what should you do *given* some goals or preferences) that is not the same as relativism.[5]

5 Another helpful way of thinking about the difference is that instrumental rationality still makes room for critiquing decisions, while standard accounts of relativism rebuff judgment.

2.5   Independence of Options and States

When we have been organizing our tables, we have been more or less assuming that the world states are "out of our control" while the options are the things that we can control by making a choice. That is, we have been assuming that options and world states are somehow independent of one another.[6] This is a common assumption that is made in decision theory. An outcome (a cell in the table) is a combination of an action (a row) and something the world does (a column). When we insist that states and options are independent, then dominance reasoning seems pretty straight forward.

6 These notions of control and independence are central and we'll return to them again later.

What if we weren't so strict about such independence? What if we don't "factorize" the outcomes into the two separate components of states and options? If we don't, then we may find that dominance reasoning isn't as straightfoward as we might think. Here is an example from Jim Joyce.[7] Imagine you park your car in a sketchy neighbourhood and someone promises to "protect" your windshield from harm. They offer their services for a mere $10. You know that those who refuse this offer tend to find their windshield smashed, while those who take it don't. You also know it costs about $400 to replace your windshield. We can represent this example of extortion with the following decision matrix, where we've put the utility number in front and the monetary cost in parentheses.

7 *The Foundations of Causal Decision Theory*, pp115-6.

	Broken Windshield	Unbroken Windshield
Pay Extortion	1 (-$410)	3 (-$10)
Don't Pay	2 (-$400)	4 ($0)

Let's use dominance reasoning. This means that to decide what to do, we check whether there is a unique option that is not dominated. When we look at the state of Broken Windshield, the Don't Pay option is better than Pay Extortion. When we look at the Unbroken Windshield state, we again see that Don't Pay is better than Pay Extortion. Because in each state Don't Pay is better than Pay Extortion, the Pay Extortion option is strongly dominated by Don't Pay. There are no other options and Don't Pay is not dominated by any other

option. So according to dominance reasoning, you should not pay the extortion.

But wait. You know very well that if you don't pay the extortion then the most likely outcome is that you'll find yourself with a broken windshield. You're not really saving yourself $10 at all, you're effectively guaranteeing that you'll have to pay $400. Between the outcome of losing $400 and paying $10, you'd rather pay the extortion. And so, you reason, you should pay the extortion.

What's gone wrong? Dominance reasoning says you shouldn't pay the extortion money, but your own line of reasoning says you should. Is the dominance principle a bad rule after all?

Part of what's going on is how we've set up the decision matrix. While the matrix sets up an outcome for not paying and having an unbroken windshield, it's less clear that this outcome is the result of two independent components. In fact, there seems to be a causal relationship between the options and the states. The state of an unbroken windshield will depend in part on whether you choose to pay the extortion or not. What we want when we set up a decision matrix is the following: outcomes should depend on the combination of a state and a choice, but states and choices should not depend on each other.

Later we will see different ways of thinking about what it means for states and choices to be independent, and some of them will lead to different recommendations. For the time being, we will stick to clearer cases where states are independent of options. In these cases, dominance reasoning seems to align with our intuitive judgments of what to do.

Here's an example where worlds states are independent of options. Suppose your friends are trying to convince you to go camping this weekend. You're indifferent - you like staying home as much as you like being with your friends when camping. Except when it rains. You really hate camping in the rain and you'd much prefer to stay home. Here's a decision matrix set up for us.

	Rain	No rain
Go camping	1	3
Stay home	2	3

Following dominance reasoning, you should stay home because that option weakly dominates the option to go camping. This setup follows the above guideline that states should be independent of your choice: whether you go camping or stay at home will not impact the weather.

Here's an example of how things can go wrong in our modeling of a decision if we don't pay attention to the guideline of keeping

options and states independent. Suppose, as we have in many of our examples, that you've got an exam tomorrow and suppose you're deciding whether to go to a party (and thereby not study properly) or to study (and thereby not party properly). It may be tempting to set up the states into Pass Exam or Not Pass Exam:

	Pass Exam	Don't Pass Exam
Party!	4 (pass and party!)	2 (fail and party!)
Study!	3 (pass but no fun)	1 (fail and no fun)

Under this poorly created setup, you might reason that you should go party, since for each state the option to party is better than to study and therefore strongly dominates the option to study. This setup, however, is poorly designed. There is a causal relationship between studying and passing the exam, which means the options and states are not independent in the above decision matrix. Again, when we set up what the states are, we should think of them as those things that are outside of our control, while the options are the things that we do control. Passing an exam is something that you have at least *some* control over, so it cannot be a candidate as a state in a decision matrix.

We can rework the example like the way we started this chapter, by understanding that the level of difficulty of the exam is up to the professor, it is not in your control. Moreover, we supposed that you're uncertain about whether your professor will create a difficult exam or not. Your choice to study or not will have an impact on the outcomes in terms of the grade you can expect to receive, but your choices are independent of the states (what questions your professor chooses). By taking this kind of care in our setup, we then have the kind of table that we started out with and that we can apply our decision rules to. The only thing that's left missing in this table is the preference assignment, i.e. the ordinal utility that you would give to each outcome to rank them from best to worst:

	Difficult Exam	Easy Exam
Party!	(C grade and party!)	(B grade and party!)
Study!	(B grade and no fun)	(A grade and no fun)

Exercises

1. Peter is considering two different lotteries. Lottery 1 costs $20 to play and there's a one in a billion chance that he could win the $1M prize. Lottery 2 costs $10 to play and there's a one in

hundred chance that he could win the $10k prize. Suppose Peter chooses to play Lottery 1. Which of the following decision rules is Peter most likely using?

a. Maximax
b. Maximin
c. Dominance (weak)
d. Dominance (strong)

2. Consider the following table where the world states (columns) are the levels of difficulty of an upcoming exam and the rows are the options you are considering on Thursday evening:

/	Very Hard	Somewhat Hard	Easy Exam
Party!	1	6	12
Study!	10	5	6
Call Family	3	4	5

- What does the weak version of dominance recommend?
- What does the strong version of dominance recommend?
- What does maximax recommend?
- What does maximin recommend?

3. Using tables like the ones in this chapter as a starting point, but changing around numbers, see if you can think of examples of the following:

- Is there an example of a preference assignment where Maximin and Dominance Reasoning make different recommendations? Why or why not?
- Is there an example of a preference assignment where Maximax and Dominance Reasoning make different recommendations? Why or why not?

4. Think of a decision you've made in the past, say like whether to go to university or not. Make a table to represent such a decision, providing what you took to be the options and states. Then plug in numbers to rank the outcomes. Then apply the decision making strategies from this chapter. What recommendations do they make? Do they align with what you decided to do?

5. What kind of decision making strategy do you tend to use? Compare and contrast your strategy with the ones we've covered in this chapter.

6. One benefit of differentiating between different kinds of decision making strategies is that it can help us better understand other people's choices. Given the same decision matrix, two people could come to different conclusions of what do simply because of differences in their decision strategies. Think of an example of how the strategies in this chapter could lead two people to come to opposite recommendations, *even if they agree on the same table*. Do some of these strategies have associations with psychological dispositions, like people you consider to be more introverted vs more extroverted?

3
Transitivity and Completeness

So far we have been using *ordinal utilities* to represent preferences, without having said much about exactly what utilities are other than we can rank them from best to worst (with possible ties). There are different accounts of what preferences are, and we'll get to some details later. All accounts of preferences, however, seem to agree on some core axioms - "obvious claims" - that govern preferences. We'll cover two in this chapter that apply to ordinal utilities. Other axioms will require us to have covered the concept of probabilities and other notions of utilities in more detail, so we'll wait to introduce those axioms later.

The two axioms we cover here are known as **transitivity** and **completeness**. Let's use ice cream flavours to briefly illustrate these. Transitivity says that if I prefer strawberry to chocolate, and I prefer chocolate to vanilla, then I prefer strawberry to vanilla. Completeness says that either I prefer strawberry to chocolate, or I prefer chocolate to strawberry, or I'm indifferent between the two flavours. The same goes when comparing these to vanilla.

Transitivity and completeness play important roles in thinking about what preferences are, particularly in the context of using ordinal utilities to rank things. In fact, they apply not only to the case of individuals making decisions, but also in thinking about how to aggregate preferences of individuals into a collective group preference. There are several ways of doing that aggregation, like voting, but we might think that some are better than others depending on how well they reflect the preferences of individuals.

The thought is, preferences are or should be transitive and complete, regardless of whether those preferences belong to an individual or to a group. The very rough reason is that decision making seems to require putting options into some kind of ordering (with possible ties), and any ordering will satisfy both the transivity and completeness axioms. The tools we have been developing won't work without being able to order things. But as we'll see, there are independent

arguments for transitivity and completeness that go beyond the convenience they provide for being able to use our decision theory tools.

3.1 Notation

It will help to have some notation to express transitivity and completeness more succinctly. Let's say that '$a \succ b$' means that a is wanted more than, or preferred to, b.[1] If instead we want to say that b is preferred to a, then we would write '$b \succ a$'. If we want to express indifference between the two, because you'd be equally happy (or sad) with one or the other, then we'll say '$a \sim b$' or '$b \sim a$'. In some cases we'll want to say that 'a is at least as good as b' and we'll use the notation '$a \succeq b$'.[2]

Here are the two axioms using this notation, using x, y, z as variables for outcomes (or objects).

Transitivity If $x \succ y$ and $y \succ z$, then $x \succ z$.

Completeness Either $x \succ y$ or $y \succ x$ or $x \sim y$.

Notice that the axioms are stated in a way that leaves them ambiguous to descriptive and normative interpretations. If the decision theory we're doing is descriptive, then these axioms describe real decision makers. If the decision theory is normative, then these axioms describe ideal decision makers - how people *ought* to approach decisions.

In the descriptive case, the transitivity axiom is false if John prefers strawberry to chocolate and chocolate to vanilla, but prefers vanilla to strawberry. In the normative case, we would say that John is not being rational.

If John claims that he neither prefers strawberry over vanilla, nor vice versa, and moreover that he's also not indifferent to them, then the completeness axiom is false under the descriptive interpretation. In other words, completeness is false if strawberry and vanilla cannot be compared. Under the normative reading, John is failing to be rational: surely he's at least indifferent to either flavour if he doesn't prefer one over the other!

This last point almost sounds like an argument for why we expect people to have preferences that obey completeness - normatively, and perhaps descriptively too. The arguments that we're going to look at are typically used for the normative interpretation of the axioms. We'll look at arguments for or against each axiom in more detail in the following sections. But first, we need to understand a particular kind of conceptual tool that is used in these arguments.

[1] The symbol '$\succ$' is just like $>$ except it applies to preferences between objects or outcomes, whereas $>$ applies to the magnitude of two quantities. But for both symbols, imagine they are mouths about to eat the preferred things.

[2] You can read this as 'a is strictly preferred to or equally as good as b'.

3.2   *Money Pump Arguments for Axioms*

We might dig in our heels and insist that the whole point of calling the above principles 'axioms' is that they do not require justification. Most decision theorists, however, think we can do better than that, i.e., we can provide arguments that justify the axioms.

Arguments for preference axioms tend to be pragmatic ones. Pragmatic arguments show that if someone violates an axiom, then they are guaranteed to lose in a decision problem - a decision problem that 'rational' agents would not accept. One of the most famous pragmatic arguments is known as the *Dutchbook Argument* which we'll get to when we cover the axioms of probability theory. When it comes to axioms of preference, the type of argument we'll see is known as a **Money Pump Argument**.[3]

Here's an example of a money pump argument. Suppose, contrary to our definition, that $a \succ b$ and $b \succ a$, which is like saying that John prefers strawberry to vanilla, and also prefers vanilla to strawberry. If you're scratching your head as to how that's even possible, good. That means you're getting the feeling of this very odd scenario, and we're going to turn that tension into an argument. Suppose that John has some strawberry ice cream and you have some vanilla ice cream. Since John prefers strawberry to vanilla, you offer John one cent to trade. John accepts. In fact this is a good deal for him since the one cent cost is negligible. Now that John has the vanilla and you have the strawberry, you again offer John to trade for one cent. John should again accept, since he also prefers vanilla to strawberry. Since John's preferences are such that $a \succ b$ and $b \succ a$, you can 'pump money' out of John over and over again by cycling back and forth between his preferences. The pragmatic argument now says that this is absurd. Surely even John should recognize his failure and give up at least one of the two preferences.[4]

The strategy behind money pump arguments is to build a sequence of trades where the agent accepts each one, but by the end of the sequence of trades they are right back where they started, except with less money. In John's case, he started with strawberry ice cream, made two trades, ending up where he started (with strawberry) but with two cents less.

3.3   *Arguments for Transitivity*

It's important to recognize that not all relations are transitive. 'Birth mother' is not transitive: If Sally is the birth mother of Laura, and Laura is the birth mother of Jill, it is not the case that Sally is the birth mother of Jill. Another example of a non-transitive relation is '*x*

[3] Some think that pragmatic arguments are a weakness and it would be preferable to justify axioms and other principles in some purely theoretical way. Others think that pragmatic arguments are actually a strength of the theory because it connects it to practice and actions.

[4] TODO: Make and insert a diagram.

is a better team than y'. It is not uncommon that team A wins a game against team B, that team B wins against team C, but also that C wins against A. On the other hand, 'is taller than' is transitive: if John is taller than Bob, and Bob is taller than Paul, then John is taller than Paul.

Decision theorists claim that preferences are transitive. That is, whenever there are two preferences, $x \succ y$ and $y \succ z$, then there is a third, $x \succ z$. If preferences are transitive, then it serves as a reason for avoiding the following money pump decision scenario.

Let's say, contrary to the transitivity axiom, that John's ice cream preferences are such that S≻C, C≻V, and V≻S (where 'S', 'C', and 'V' stand for strawberry, chocolate, and vanilla, respectively). We're going to set up a sequence of trades that John should accept, but land him in the same starting position, except with less money. Say John starts with chocolate. Then since S≻C, he'll accept a trade for strawberry for a very small price, say one cent. Now with strawberry in hand, John would also accept a trade for vanilla, especially if this trade is free. Finally, John should also accept a trade back for chocolate, especially if, again, it's free. Now John is back in the original starting place (with chocolate), except with one cent less. We could also have charged John a small for the other two trades, or for any one of them - it doesn't matter so long as some cost, no matter how small, is charged somewhere along the way. Since the sequence of trades takes us back to the same starting condition, we can repeat the sequence again, and thereby 'pump' money out of John indefinitely.[5]

[5] TODO Make and insert a diagram.

Compare Sally to John. She also holds that $S \succ C$ and C≻V, but contrary to John, her preferences are consistent with the transitivity axiom and she holds that S≻V. So, like John, she would accept a trade for chocolate if she had vanilla, and she would also accept a trade for strawberry if she had chocolate. But unlike John, she would not trade for vanilla if she had strawberry, because that is the opposite of her preferences given that they are transitive. Sally thereby avoids the money pump scenario. Because Sally's preferences are transitive, she breaks the cycle that John cannot.

The transitivity axiom for decision theory claims that preferences are transitive. What it says is that some ordering of preferences are to be excluded, namely those that are not transitive. It is this ability to exclude non-transitive preference orderings that provides a reason for avoiding the type of money-pump scenario just described. A non-transitive preference ordering either has no reason for avoiding the scenario (which seems bad) or at the very least has the burden to provide some other reason. A person like John might foresee how they're going to be money pumped and decide not to accept any more trades. But that alone seems like a poor reason, since it

acknowledges the badness of the consequences of non-transitive preferences, without suggesting an alternative. The burden falls on them to provide an explanation of how they can avoid being money pumped. That's why money pump arguments are pragmatic - it is a kind of burden shift, rather than a proof that no other reason exists.

3.4   *Arguments for Completeness*

Our ice cream example makes completeness seem almost trivial: either strawberry is preferred to chocolate, or chocolate is preferred to strawberry, or one is indifferent between the two. But the completeness axiom doesn't just hold for ice cream flavours, it holds for all objects and options that we can have preferences over.

It is often said that you can't put a price on a human life. In our way of putting it, that means for any given dollar amount x and for any life l, it is neither the case that $x \succ l$, nor $l \succ x$, nor $x \sim l$. This claim is saying that money and lives are not comparable. That is, they are *incommensurable*.[6]

The completeness axiom says that you *can* compare money and lives. We might not like it, but the completeness axiom says it can be done. If government decision making is any indication, the completeness axiom is probably correct, though no politician would probably admit it. Here's a brief reason to think that governments do, at least implicitly, compare lives and money. If a citizen is lost at sea, for example, we may be willing to pay out a lot of money to organize a search and rescue party. Surely, however, there is a limit to how much money we would be willing to put up. Maybe we'd pay a million, or even more, but surely there is an upper bound, some value where, though it pains us deeply to have to admit it, the life is not worth the inordinate cost. While this example may be extreme, more route examples abound. Decreasing speed limits on highways would save lives, but people save time by driving faster, and that time can be used for work or leisure - and either way we can put a dollar amount on that. So the decision of where to set a speed limit is implicitly comparing lives to other things, like time and money.

Examples aside, there is a money pump argument for the completeness axiom. Not everyone finds it as compelling as the money pump argument for transitivity, but nevertheless, here it is. Suppose there are two objects, x and y that are said to be incommensurable. In addition, suppose that y^+ is just like y but is strictly preferred to it.[7] Because $y^+ \succ y$, a small cost would be acceptable to trade y in order to get y^+. If we accept that it is permissible to swap incommensurable objects, then the money pump argument is up and running: start with x, swap for y, now pay to get y^+, then swap back for x,

[6] Arguably there are word meanings that are incommensurable across languages. For example, some have argued that there is no translation in English of the German word 'Weltanschauung'.

[7] Think of + like a small bonus.

which is where we started.[8]

This money pump argument isn't quite as compelling because it relies on the assumption that incommensurable objects can be swapped. One might try to argue that incommensurability is supposed to be conceptually distinct from indifference. If we are indifferent between two objects, then it seems perfectly fine to swap between them. But the main thrust of saying that two objects are incommensurable is to say that is not straightforwardly permissible to swap (note this is not quite the same thing as saying that it is impermissible).

Putting aside the question about the conceptual difference between incommensurability and indifference, we can ask how it might show up in behavior. When you make a choice, we take this to "reveal" or provide evidence of your preferences. For example, suppose a subject is presented with objects a and b, and they choose a. This is evidence that they prefer a over b, i.e., $a \succ b$. (I'm cheating here by letting a and b refer to both the objects and the corresponding options, but not much is to be gained by being pedantic.) Of course mitigating circumstances might explain why they chose a even though they prefer b: perhaps a would give them an opportunity to still choose b later (but not vice versa), or they were threatened (unbeknown to us) or something else. So it is a tricky business to infer preferences from choices that people make, but in some idealized circumstances, there is at least a rough inference we can make that goes from behavior to preference (more on this when we analyze utilities in more detail). How could we ever tell the difference, however, between incommensurability and indifference? The objection, from this perspective, is that any behavioral test for incommensurability can also be used for detecting indifference. We might for example, observe that sometimes a is chosen - in which case we would say that a is at least as good as b ($a \succeq b$) - and other times b is chosen - in which case b is at least as good as a ($b \succeq a$). Since each object is at least as good as the other, we are invited to infer $a \sim b$. It is unclear how to design a behavioral test that would detect incommensurability without inviting the inference to indifference.

Another way to handle the possibility that some objects are incommensurable could be to restrict the domain of decision theory. If some objects are incommensurable, then those aren't the kinds of things we can have preferences over, and the theory can't make decisions about those things. This response somewhat captures the difficulty raised by the claim that, e.g., you can't put a dollar amount on a human life. One can understand this sentiment as trying to express that we can't, or ought not to be engaging in decisions when these objects are being considered.[9]

The issues we've raised about completeness assume that comparisons must be done on a single scale. One strategy for handling these difficulties is to allow for different scales and then find ways of comparing the scales. This is known as the multi-attribute approach. However, to understand this approach requires a more sophisticated understanding of what sorts of things utilities are, and we'll cover that in a following chapter.

For now, we're going to look at other applications of decision making when we assume that preferences are transitive and complete. More specifically, we'll look at how groups can make decisions given preferences of individuals.

3.5   Social Choice

3.5.1   Group decisions are not mere sets of individual decisions

Almost all of the examples we have covered have been about individual choices. In many cases, however, it is not an individual that makes a choice, but a group of individuals. This leads to a separate set of questions about how groups do/should make choices. At the very least, the way a group makes a choice should in some way reflect the preferences of the individuals. It would be extremely odd if everyone unanimously agreed that the best option for the group is A, but then some other option is picked "by the group'' that is unanimously agreed on to be the worst option.

That said, we will see that there are different ways in which we can be explicit about how a group (a collective of indivbiduals) should be connected to the preferences of the individuals that make up the collective. Social choice theory is about how to aggregate preferences of individuals into a single collective preference. It turns out that there are many ways of doing this aggregation. However, there can be tensions between the individual level and the collective level, particularly concerning the notion of rationality.

Here's a brief illustration of how such a tension could come about. Suppose that there are three judges, J1, J2, and J3. Each judge is "rational", which means at the very least that they subscribe to a rule of inference philosophers call **modus ponens** (a latin name for affirming the antecedent). Modus ponens says that if you believe a conditional like 'If A then B' is true and you also believe that 'A' is true, then you should infer that 'B' is true. For example, suppose you know that if Johnny eats peanuts then he will have an allergic reaction, and you also know that Johnny is eating peanuts, then you should infer that Johnny will have an allergic reaction![10]

Now suppose that each of our judges has been called on to make

[10] Note that the other direction isn't valid, i.e., if you know that Johnny is having an allergic reaction, it's not guaranteed that that's because Johnny ate peanuts - he might have eaten something else that he's allergic to.

a decision about a case. The case is concerned about a purported incidence of theft of a water bottle. The judges are asked to assess the truth of the following three claims:

S1 The defendent stole the water bottle.

S2 If S1, then the defendant should go to jail.

S3 The defendant should go to jail.[11]

Each judge makes their own decision about the case, which are then collected in the table below.

[11] Notice that if S1 and S2 are true, by a matter of logic S3 is forced to be true too.

Judge	S1	S2	S3
J1	True	False	False
J2	False	True	False
J3	True	True	True

Notice that each judge meets the rationality constraint of obeying modus ponens: the only judge that says S3 (the defendant should go to jail) is the judge that also says S1 and S2 are true. In order to say that S3 is false, at least one of S1 or S2 would have to be false, and indeed that is consistent with what Judges J1 and J2 say.

In other words, each judge is individually rational. But what about the collective? For example, what if we aggregate the judgments of the judges and use the idea of majority to assign truth or falsity to the three statements? In other words, we might use **majority ruling** to aggregate the decisions into a group decision. If we do that, then we would have:

Judge	S1	S2	S3
J1	True	False	False
J2	False	True	False
J3	True	True	True
Group	True	True	False

Notice that the collective judgment *violates* modus ponens, that is, it says that S1 and S2 are true, but S3 is false. Aggregating judgments by majority voting does not, in other words, preserve rationality from the individual level to the collective level.

Rationality is not the only thing that we might be concerned about when we aggregate things from the individual to the collective level. Not to mention, rationality is somewhat complex even at the individual level. So let's turn to something simpler.

What we'll focus on for the remainder of this chapter is whether there are ways to aggregate preferences in a consistently acceptable way. What counts as ''acceptable'' is something we'll discuss.

To start, we'll look at some examples of how preferences might be aggregated, keeping track of advantages and disadvantages.

3.5.2   *Plurality and Runoffs*

One common form of aggregation, especially in elections in the United States, is plurality voting, also known as 'first past the post' or 'winner takes all'. Let's look at it in a simple context. Suppose there are seven people, P1, P2, . . . , P7, that are long time friends and are deciding on where to go to dinner together. There are four restaurants they are considering, R1, R2, R3, and R4. Each person gets one vote. Suppose the table below summarizes the results of their vote.

	P1	P2	P3	P4	P5	P6	P7	Total
R1	Yes	Yes	Yes					3
R2				Yes	Yes			2
R3						Yes		1
R4							Yes	1

By these results, it looks like R1 should be the choice of the group, because it has the most 'plurality' of votes.

Note, however, that while R1 has the most number of votes, it does not have the majority of votes. Suppose the P4-P7 are all strongly opposed to going to R1. Then an odd consequence of plurality voting is that that group goes to a restaurant that the majority of people in the group oppose.

One way to handle this is to have a 'runoff' vote. If no option, such as an election candidate or restaurant (in our case), is the majority, then we hold a second vote between the top two vote getters. In our case that is R1 and R2. Suppose that such a second vote is held and we get the following results.

	P1	P2	P3	P4	P5	P6	P7	Total
R1	Yes	Yes	Yes					3
R2				Yes	Yes	Yes	Yes	4

In a runoff vote the recommendation we get for the group is to go to R2.

What plurality and runoff voting have in common is that they are aggregating people's first preferences. Little to no weight is given to people's second, third, or even last preferences.

This has some advantages. We saw above money pump arguments for thinking that individual preferences should be transitive and

complete. One might like this to also be true of group preferences. So long as we restrict ourselves to the combination of plurality and runoff voting, the group preference will also be transitive and complete. But the reason for this is because the plurality+runoff system is ultimately just aggregating preferences for two options. Either there is an option that gets a majority of first preferences or not. If an option a gets a majority of first preferences, then a received at least one more than half the votes. Then for any other candidate x, they can have at most one less than half the votes. So, a will be preferred to x and it's not possible to violate either transitivity or completeness. If an option a does not get a majority of first preferences, then it might be possible that a has more votes than another or less. But if a is one of the two options with the highest votes, then the runoff vote forces another vote between those to options, say a and b. In this runoff vote, either a will get more than half, exactly half, or less than half of all the votes, which correspond respectively to the group preference being $a \succ b$, $a \sim b$, or $b \succ a$. In each scenario, neither transitivity nor completeness are violated.

But a system that effectively ignores everything but people's "first" preference seems like an impoverished system. The consensus we get in the plurality+runoff system is forced. To see this, consider another system that takes into account other preferences besides the first. This is called the *Borda count*.

3.5.3   *The Borda Count and an Impossible Task*

We'll illustrate the **Borda count** with our restaurant example. Let's suppose that each person ranks the restaurants, so that 4 points goes to the most preferred restaurant, 3 to the second, 2 to the third, and 1 to the least preferred. If we had five restaurants, we would have 5 points go the first, 4 to the second, etc. In general, if there are n restaurants, we give n points to the most preferred, $n - 1$ to the second, $n - 2$ to the third, etc. This method of aggregation is used in some college sports in ranking teams. Notice that it doesn't just take into account people's first preferences, but all preferences. Also notice that it does not capture how much more you might like one option over another. Again, all we are doing here is putting the outcomes in an order using ordinal utilities (an ordering that, at the individual level, will be transitive and complete).

Suppose we get the following results in our restaurant example.

	P1	P2	P3	P4	P5	P6	P7	Total
R1	4	4	4	1	1	1	1	16
R2	1	3	3	4	4	2	2	19

	P1	P2	P3	P4	P5	P6	P7	Total
R3	3	2	2	3	3	4	3	20
R4	2	1	2	2	2	3	4	15

Notice that the first preferences still align with our previous tables (all the 4's are in the 'yes' positions). Also notice that R1 is still the least preferred for P4-7, which means we preserve the result we got from the runoff vote. However, unlike the recommendations of the plurality vote and the runoff vote, the Borda count recommends that the group go to restaurant R3.

So the Borda count has a nice feature in that it builds consensus by considering people's preferences beyond their first. Even though the group goes to a restaurant that is almost no one's first preference (in our example R3 does happen to be P6's most preferred), it finds a restaurant that is broadly acceptable in the group.

So far so good, but we might now ask: is the Borda count a way of aggregating individual preferences so that, if the individual preferences are transitive and complete, the group preference will be transitive and complete as well?

Let's look at the following example. It will help to organize our table a little differently so that we can tell just by looking what the vote tallies would be. Rather than put options in the rows, we'll organize the rows in terms of people's most to least preferred options, going from the top to the bottom. Let's say we have three people, P1, P2, and P3, and they are deciding between three books to read together next month: A, B, or C. The table below summarizes the three people's preferences from best (1st) to worst (3rd).

Preference	P1	P2	P3
1st	A	B	C
2nd	B	C	A
3rd	C	A	B

If we just compare the preferences between books A and B, then A would get 2 out of the 3 votes (whatever those points would be). If we just compare B and C, then B gets 2 out of 3 votes (2/3). So this might suggest that, in terms of the group preference, A is preferred to B and B is preferred to C. But when we inspect further and compare A and C, we'll see that C gets 2/3 votes. Our pairwise comparison produces a cyclical group preference: $A \succ B$, $B \succ C$, and $C \succ A$. This is in violation of the constraints that group preferences should be transitive, i.e., if $A \succ B$ and $B \succ C$, then $A \succ C$, not $C \succ A$!

One potential response to this cyclical situation is to say that the group should be indifferent to A, B, and C, i.e., $A \sim B \sim C$. The problem with this suggestion is that we can't both accept it and the condition called **the independence of irrelevant alternatives** (or IIA). Suppose, for example, that book C is on back order and will not be delivered in time for the group to read next month. Book C is now an "irrelevant alternative" and we should drop it. But here is the problem. If we agreed that the group should be indifferent between the three groups, $A \sim B \sim C$, then dropping C, the irrelevant alternative, should not affect whatever preference holds between A and B. That is, after dropping C, the group should still be indifferent between A and B, i.e. $A \sim B$. The preference between A and B is said to be *independent* of the *irrelevant alternative*. But let's look at the table again if we drop C there:

Preference	P1	P2	P3
1st	A	B	A
2nd	B	A	B

The table makes it pretty clear that the group should prefer A to B. But maybe the group decides to hold themselves to their original agreement and uphold IIA. That is, since they had agreed that $A \sim B \sim C$ before they knew that C was on back order, the group preference should still be $A \sim B$ - despite what the new table says. The problem with this proposal is that it does not maintain the implicit fairness that each person's preferences gets equal weight in the group. To be indifferent between A and B, the new table would have to assume that P2's preference for B is worth twice as much as P1's preference for A (and likewise P3's). That goes against the very point of aggregating their preferences through something like the Borda count. P2 would be something like a "dictator" in shaping the group preference.

Maybe another strategy is to leave behind the Borda count method entirely and instead require a consensus - a unanimous vote - in order for the group to prefer one option to another. If there is no consensus among the voters between two options, then the group shall be indifferent between those options. Looking at our tables, it is clear that there is no consensus that A is preferred to B, since P2 prefers B to A. This 'method of consensus' is nice in that it satisfies IIA (independence of irrelevant alternatives) and does not give more weight to one person's vote over others. In fact, this 'method of consensus' seems like the only possible solution to the 'cyclical preferences' issue.

The problem with the consensus method suggestion, however, is that it violates transitivity. Suppose instead of the summary tables above, we had these results instead.

Preference	P1	P2	P3
1st	A	B	C
2nd	B	C	A
3rd	C	A	B

The group unanimously votes that A is preferred to C, and so by consensus we get $A \succ C$. If we compare A and B, we notice that there is no consensus, and so by this method the group should be indifferent, i.e., $A \sim B$. Similar inspection of the table means $B \sim C$. Since we have $A \sim B$ and $B \sim C$, then by the transitivity constraint we should have $A \sim C$, but our consensus method told us $A \succ C$!

The tension we are feeling is actually a very deep problem. It can be laid out explicitly in a formal argument and is known as **Arrow's Impossibility Theorem**. What it says, put intuitively, is that for any voting system that allows for three or more options, at least one of these three things will *not* hold:

1. Unanimity: If everyone prefers A to B, then the group prefers A to B.
2. Independence of Irrelevant Alternatives: If nobody changes their mind about the relative ordering of A and B, then the group can't change its mind about the relative ordering of A and B.
3. No Dictators: For each voter, it is possible that the group's ranking will be different to their ranking.

To be clear, we are not saying that all voting systems are bad: some are worse than others. Presumably a voting system that satisfied none of these conditions is pretty unacceptable, especially in comparison to a voting system that at least satisfies No Dictators and Unanimity. What Arrow's Impossibility Theorem tells us is that it's impossible to have it all, i.e. a voting system that obeys all three features.

3.6   *Limitations and Key Take Aways*

All that said, we have narrowed in on a result about group preferences that is deep, but not without a fair number of substantial background assumptions (some of which we have left implicit). For many group decisions, like friends choosing what restaurant to go to, some of the assumptions either might not hold, or the conditions

don't impose a level of unfairness that we care too much about. For example, while we might have to allow one friend to dictate what restaurant to go to this Friday night, we might allow another friend to dictate where to go the following week. Or we might allow the group to split and go to do different restaurants (whereas in elections it is not typically possible to elect two presidents simultaneously). In any case, the range of Arrow's Impossibility Theorem is limited to several background assumptions, and while many political elections meet those assumptions and thereby face certain impossibilities, lots of group decisions don't meet those same background assumptions.

So we have seen the kinds of uses that ordinal utilities provide us with, particularly as a tool for ranking things, both on the individual and group level. They have also helped us introduce important concepts like transitivity and completeness, which we will make further use of in later chapters. But they have important limitations. Two are particularly relevant.

First, while ordinal utilities allow us to order outcomes from best to worst, they don't provide us with a way to say *how much more* we might like one outcome over another. In our restaurant example, each person ordered or ranked their restaurants from most preferred to least preferred. But it possible that the difference between one person's first and second choice is very small, while for another person it is very large.

Second, ordinal utilities don't give us a way to compare how much one person's first choice is liked in comparison to another's: one person might absolutely love their first choice, while another person sees their first choice as bad, but the least bad among all the options. We'll see whether a different kind of numbering system that uses *cardinal* utilities will help us with these two limitations. (Spoiler: they help for one, but not the other.)

Exercises

1. Transitivity is a kind of pattern that some relations have. Intuitively, a relation can be anything that involves two objects. Examples include "John is left of Sally", "this stone is heavier than this chair", and "Oliver is funnier than Connan". It's helpful to think of transitivity in terms of a pattern that a relation R might have: if xRy and yRz, then xRz. Think of different kinds of relations you could plug into R and the kinds of objects that x, y, and z are related by R.

   • What are examples of relations that are transitive?
   • What are examples of relations that are not transitive?

2. Similarly to above, completeness is a kind of pattern that some relations have: either xRy or yRx.

   • What are examples of relations that are complete?
   • What are examples of relations that are not complete?

   MONEY PUMP ARGUMENTS. Suppose you have two roommates named Peter and Sally. The three of you watch an episode from a series together every night of the week except Sundays. There are three different series that the three of you have been following together: Show A, Show B, and Show C. Because you own the device and pay for the streaming subscription, you get to pick which show to watch on any given night. Peter claims that he always prefers Show A over Show B, that he always prefers Show C over Show A, and that he always prefers Show B over Show C. Sally, on the other hand, is always indifferent between Show B and Show C, but she prefers Show A over either those two. Your preferences, however, change on a day to day basis, but you prefer not to watch the same show two days in a row and you want to see at least one episode from each show in any given week. That said, let's suppose that you'd be willing to let one of your roommates pick the show in exchange for one dollar. Explain how you could "money pump" one of these roommates but not the other. In your explanation, feel free to imagine that you'll make offers to Peter on every day for one week, and then to Sally on every day another week. What sequence of selections in a week would guarantee that you'll get $6 from exactly one of those roommates?

3. Make a table like the Borda count restaurant example, but so that the group choice is not anyone's most preferred choice.

5. Make a table like the Borda count restaurant example, but so that the group choice is at least one person's least preferred choice.

4
Utilities

There are three broad questions we would like to answer in this chapter.

1. Is there a way for a person to represent *how much more* one option is preferred to another?

2. Is there a way to compare one person's preferences to another that isn't just the ordering of the options? That is, can we say something more substantial about two people's first choice, like that one person likes it the most but another dislikes it the least?

3. How do we interpret what the numbers mean? What are the numbers referring to in the real world? That is, what is their content?

THESE QUESTIONS ARE NOT JUST PHILOSOPHICAL. There are important practical considerations at stake. For example, the World Health Organization (WHO) has to somehow measure the value of medical interventions. The unit they use in their assessments is called a QALY - quality-adjusted life-year. A QALY incorporates both the quality and quantity of life lived. One QALY amounts to living one year in perfect health, while zero QALY amounts to death. The QALY is not without controversy and the debate ranges from purely philosophical considerations to the deeply pragmatic constraints that policy-makers face.[1]

We will introduce the idea of a **cardinal utility**.[2] The basic idea is to use what's called an *interval* scale. In addition to being able to order things using numbers, the numbers are also meant to represent *magnitudes*. Money is an example of an interval scale, to some approximation. We can use money not only to order things from most expensive to least expensive, but we can also say how much more expensive one thing is to another. One might think, however, that money is ultimately limited in how fine-grained it can be - perhaps it doesn't make much sense to speak of quantities that are smaller

[1] Another example that we will look at below is how the concept of utility can motivate arguments concerning social welfare and taxation.

[2] Supposedly the first use of "cardinal utility" was in 1934 by John Hicks and Roy Allen.

than a cent. Whether or not that's the case, cardinal utilities can be arbitrarily precise and are typically represented using real numbers (as opposed to whole numbers).[3]

IF WE COULD BUILD AN INTERVAL SCALE FOR PREFERENCES, we might be able to solve some of the limitations of ordinal utilities. For example, suppose there are four books, A, B, C, and D that both Alice and Bob are considering to read together next month. Using just ordinal utilities, it is easy enough for Alice and Bob to order the books with respect to their preferences: each will use 4 to represent their most preferred book, 3 the next preferred, and so on. But remember that we allow for ties, say between B and C. That is, it could be the case that both Alice and Bob agree in their preference orderings, so that $A \succ B \sim C \succ D$. But what should we make of the situation where Alice assigns 4,3,3,1 for books A, B, C, D, respectively, while Bob on the other hand assigns 4,2,2,1 to them? These numbers preserve the same ordering, and so from an ordinal utility perspective they are the same, but Alice and Bob seem to want to say something different about how B and C compare to the extreme ends of their best and worst options. How do we even go about this?

4.1   Creating an Interval Scale

THE LOTTERY OPTION PROCEDURE is one way to capture information about how much one option might be preferred to another. Suppose we have just three options and Alice orders them like this: $A \succ B \succ C$. We want to find how much more Alice prefers B to C, and how much less she prefers B to A. To do so, we create a new option, L. We can think of option L like a lottery that will have A and C as "prizes". But what should the odds be of winning C in lottery L compared to the odds of winning A? Suppose lottery L has 100 tickets. Given that Alice prefers A over C, it seems that A should be on all of the tickets. But what we want to find out is how B compares to A and C. So what we do is figure out at what point Alice would be indifferent between B and L, where L specifies the fraction of all tickets where A wins and the fraction where C wins. Let's see some examples to illustrate.

Let's say lottery L has 50% tickets with A and 50% tickets with C. We ask Alice which of the preferences she has: $B \succ L$, $L \succ B$, or $B \sim L$?

- If Alice says $B \succ L$, then this means that she prefers getting B to a 50% chance of getting A. That is, she prefers B more than "half

[3] The "real" in "real numbers" doesn't mean something like "actual numbers." The history of numbers and number systems is fascinating (see the Wikipedia entry for "number"). The main thing for our purposes is just to know the difference between integer numbers and real numbers.

as much" as A. But that still doesn't tell us how much more than "half as much" as A. So, in this case we would change the lottery L to a new lottery L' by *increasing* the fraction of tickets that would go to A and then ask for her preference again, comparing B to L'.

- If instead Alice had said $L \succ B$, then that means that she prefers a 50% chance of getting A than getting B. Here we would change the lottery L to a new lottery L' by *decreasing* the fraction of tickets for A.

- (Indifference) Finally, if instead Alice had said $B \sim L$, then that means she is indifferent between getting B and a 50% of getting A. Here we wouldn't have to change anything about the lottery.

Suppose we have gone through this process and we find the lottery where Alice's preference between B and L is indifferent, i.e. $B \sim L$. And let's suppose that L has 75% of the tickets go to A and 25% to C. The idea behind creating an interval scale from this information is to think that Alice's desirability of B is 75% of whatever her desirability of A is. For example, let's suppose that Alice's utility of A is 1 and her utility of C is 0. Then her utility of B would be 0.75. If instead Alice's utility scale were 0 to 8, with C at 0 and A at 8, then her utility of B would be 6.[4]

So the introduction of a lottery option helps us make important progress on determining the magnitudes of utilities, but only *relative* to the utilities of other options. That is, the lottery option procedure doesn't help us decide what the scale itself should be. It just helps us assign utilities given that we have some scale. To some extent this is a nice feature. Notice that, no matter what scale you pick, the lottery option procedure will always preserve the ordering given by the ordinal utilities, and moreover the resulting cardinal utilities will also preserve transitivity and completeness. We thus have an answer to our first question for how to represent the magnitude that one option is preferred to another.

There are, however, at least two limitations.

FIRST, because the lottery option procedure starts with ordinal utilities, which can only order options relative to one another, it cannot determine an *absolute* scale, only a relative one. In order to get an absolute scale, we would have to make additional assumptions, like that utilities start at 0 and are maximally 100. But without such additional assumptions, cardinal utilities on the interval scale will be relative to each other.

THE SECOND DRAWBACK is that it is not possible to do *interpersonal comparisons* with either ordinal or cardinal utilities, without the aid

[4] Exercise: draw this as a diagram.

of additional assumptions. To see this, note the example we had above, where Alice assigns 4,3,3,1 for books A, B, C, D, respectively, while Bob assigns 4,2,2,1 to them. Suppose we do the lottery option procedure on Alice and she is indifferent between B and L, where L has 75% of tickets go to A and 25% of tickets go to D. (Notice since Alice is indifferent between B and C, we could also have used C instead.) Then if Alice's utilities were cardinal utilities instead of just ordinal utilities, and her scale went from 0 to 4, her preference of B would be 75% of A. Since A has utility 4, then B has utiliy 3. Now suppose we do the same for Bob, and he too accepts the same lottery. Then even though he had assigned 2 to B in his ordering, we know that he prefers B the same as 75% of A. Since A had 4, then as a cardinal utility Bob would give B a 3. However, nothing about this reasoning tells us anything about what the number 4 corresponds to in Alice and Bob. As we had said before regarding ordinal utilities, Alice's most preferred option might correspond to a great deal of well-being and satisfaction, but that same option might be utterly miserable for Bob, it's just that the alternatives are even worse. For all we know, Bob's assignment of 4 is Alice's assignment of 0. Nothing about the lottery option procedure allows for a comparison between one person's scale and another's.

So unfortunately, as far as we have come to this point, the answer to our second question is negative: neither ordinal utilities nor cardinal utilities allow us to make comparisons between people's preferences beyond their relative orderings. But maybe there are principled ways of making extra assumptions that would help us justify one scale over others, and then combine it with the lottery option procedure so that we can do interpersonal comparisons. In the sections below we will look at some candidate views, which are typically connected to views about well-being or welfare. To spoil things, none of the views except for what is known as the preference-based theory tends to be widely accepted, and even it has limitations.[5]

4.2   What do the numbers mean?

4.2.1   Material and money

A naive account of well-being or value might start with material goods. Of course we all need shelter, food, water, and basic commodities to live. But it is widely accepted that material by itself cannot be a good measure of value, let alone utilities. For one, there is far too much diversity between which goods some people find valuable compared to others. I might collect whisky, while someone

[5] There is still a lingering issue that we have not made explicit. Notice that in the lottery option procedure we ask people to have a preference over an *expectation*. A lottery L is quite a different sort of thing than the options themselves. It asks us to have a preference about a probability that we get an option. This is an important complication we have introduced, and we take it up in the next chapter.

else collects dolls. If we are going to try to build some interpersonal scale, material goods will be of little help.

Money, on the other hand, can be a very useful guide in approximating the utility of an outcome. It is generally something that we all want, and something we are all willing to give in exchange for the material goods we prefer. Moreover, money also has the feature of putting things in an ordering. If *A* has more utility than *B*, then under some circumstances I'd be willing to pay more for *A* than I would for *B*. That said, there is a massive difference between the idea that money can roughly track utilities and the idea that money and utilities are one and the same. Anyone that thinks money and utilities are identical immediately faces several puzzles.

Here's one: the more money you have, the less 'valuable' an extra $1000 will be to you. Someone that earns $20K a year will find $1K more useful than someone earning $1M a year. Same amount of money ($1K), but different amount of utility.

This is a nice illustration of a more general point about the nature of money. The utility value of $2x is not twice the amount of $x. When you get the first $x, the second $x will be less valuable than the first. Economists call this feature of money *declining marginal utility*.

Here's a fictional but illustrative example. Let's suppose that the increase you get in utility for every extra dollar can be represented by: Utilities = Money$^{1/2}$. If we plot this function, we get the graph below. Notice that the first few dollars have a sharp rise in utility, but then as we continue to increase the dollar amounts the curve becomes less steep. In other words, the more money you get, the slower the amount of utility grows. The marginal utility is the utility that you get for an extra unit of good (money in this case), and it's *declining* because the size of the increase in utility is going down (the precise amount is the slope of the curve at any point, which is the derivative of our function, which is $\frac{1}{2\text{money}}$).

THE FACT THAT MARGINAL UTILITIES DECLINE means that money and utilities don't have a one-to-one relationship.[6] And because money and utilities don't even have a one-to-one relationship, they can't be identified as the same thing, even if there seems to be some kind of indirect relationship between them.

THIS INDIRECT RELATIONSHIP BETWEEN MONEY AND UTILITIES can be used to justify certain taxation policies.[7] Taxation is a way to help support society at large by funding projects that are beneficial to its members on aggregate (think roads, sewage, public spaces, etc). Except in very special circumstances, here's a bad way to fund projects: have a flat fee that each person pays each year, regardless
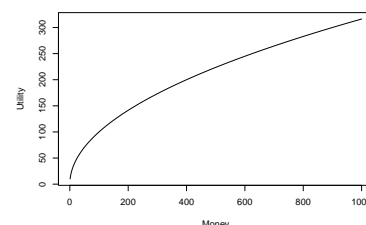

Figure 4.1: Notice that the first 200 dollars get a rise in utilities that's about 140, but the next 200 dollars (from 200 to 400) only get a rise of about 60 utilities.

[6] If it were a one-to-one relationship, we would have a straight diagonal line in our plot instead of a curved one.

[7] The history of taxation is a complicated one and the point to be made here is meant to be only illustrative, so we'll ignore numerous complications.

of their income. Suppose the flat fee is $1K per year. Superficially, it seems that this is fair, since both the person earning $10,000 (call him Tim) and the person earning $1,000,000 (call her Molly) are contributing the same dollar amounts through taxation. But a utility perspective shows that such taxation is not fair at all. The $1K coming from Tim means he has $9,000 left after taxes, while Molly has $999,000. If we look at their respective *loss* of utilities from the decreases in their spendable income, Tim is giving up more than Molly. To see this really clearly, imagine someone who only makes $1,000 (or no money at all) - they would have nothing after the flat tax fee (or even be in debt because of it!). Given these kinds of considerations, we can see why there are plausible reasons for taxing the rich "more" than the poor.[8]

[8] We put 'more' in scare quotes because, while it is more from a monetary perspective, that may not hold from a utility perspective.

THE RELATIONSHIP BETWEEN MONEY AND UTILITIES is actually far more curious than what declining marginal utilities suggest. We just saw how getting more money doesn't lead to an equivalent amount of utility. There are also examples where utility seems to be undermined by the mere introduction of the possibility of representing utilities with money.

For example, suppose a friend asks you for help to move apartments. You consider this to be a favour, one that you expect your friend to return when you move. We can see this as a kind of 'tit-for-tat' exchange of utilities: the payoff you get for helping is the promise of getting help in the future (not to mention all the warm fuzzy feelings that come with helping another person generally, let alone just common decency).

But now suppose that your friend offers you one dollar for every hour that you help. From one perspective, that should sound like a great deal: you get all the benefits we just mentioned above, plus an extra two dollars for every hour of your time!

Most people, however, do not see it this way. Once your friend offers you the $2/hr rate, you're likely to start thinking about what your time is worth in monetary terms, and your friend's offer is likely to be insulting to you. So, where from a naive perspective we might expect that the offer of money increases your preferences for the outcomes related to helping, it can actually have exactly the opposite effect. Somehow, for whatever reason, the introduction of money, even just the suggestion of it, can be parasitic on utilities related to social norms.

HERE'S AN INTERESTING TWIST. Imagine instead of money, your friend offered you beer instead (or whatever your beverage of choice may be). As long as your friend doesn't mention that the monetary

value of the beer is equal to the $2/hr rate, it's likely that you view the offer as more than fair. But if the beer is playing the role of monetary exchange, then you'd probably ask for more if you think your hourly rate is more than $2.

If your intuitions about this example are in line with this narrative, then it's consistent with empirical findings concerning favours, gift giving and monetary exchange. (CITE ARIELY) As long as money is not mentioned, social norms dictate some roughly fair exchange of social utilities: I help you knowing that in the future there's a good chance I can count on you to reciprocate. But once money gets introduced, we tend to switch over to a very different way of thinking about the exchange.

These examples are not meant to suggest that there is no relationship between money and utilities. The point is that, at the very least, there is no straightforward way to substitute between money and utilities. So with that in mind, we turn now to consider what utilities might be.

4.2.2 Hedonism and Experience-based Theories

WHAT MAKES SOMETHING GOOD? Or what makes it bad? Jeremy Bentham identified the good with pleasure and the bad with pain. On this view, we tend to avoid painful experiences because pain is identical to badness, and we tend towards good experiences because the good and pleasure are one in the same. This view gives us a neat little picture: someone's welfare is high if they have lots of good experiences and few bad ones, and welfare decreases as the number of good/pleasurable experiences go down while the number of bad/painful experiences go up.

It doesn't take long to realize that this picture gets complicated quite quickly. That is, we can find ways in which changes in pleasurable or painful experiences don't neatly track changes in welfare. For example, earning money can be a painful experience, quite literally if your job involves hard manual labour. And yet at the same time, that person might say that their welfare is going up. They may even say that their welfare is going up *because* they have to endure some painful experiences.

We might say that's not quite right. After all, it's not difficult to imagine that someone might be able to increase their welfare even further if their job were less painful. Then again, if the job were less painful, the pay may go down accordingly. At the very least, what we can say thus far is that there is not some simple relationship between pleasurable/painful experiences and welfare - we can imagine wel-

fare go up even with the introduction of some painful experiences.

There is an argument that, even if we introduce a host of complications, there simply is no account to be given that definitively connects welfare to the level of pain or pleasure of our experiences. The argument comes from a thought experiment from Robert Nozick called *the experience machine*. Here's a version of this argument.[9] We are asked to imagine a matrix-like situation in which a machine provides our brain with all the experiences of a good life - whatever that might mean to you. Like in the matrix, these experiences are not just sensations. The experiences are just like the ones we would have in real life, including the relationships we have with friends and family. The only difference is that the interactions weren't with real people. But we don't know that while in the experience machine. To the best of our knowledge, we can't tell the difference whether our friends and family actually died and they're just being simulated perfectly, or whether our interactions are with real individuals.

[9] Nozick's thought experiment was written in 1974, before the movie *The Matrix* came out in 1999. But if you've seen the movie you can easily imagine the sort of scenario that Cypher discusses.

NOZICK ARGUES that a person that's lived a life in the experience machine has lead a bad life. Why? Because their entire world has been based on an illusion. While it is true that the person in the experience machine doesn't know the difference, if we were to take them out of it, it is likely that they would agree with us that the illusory nature of the experiences they had undermines the level of welfare they would assign to those experiences. If you think that welfare and good experiences track each other perfectly, then you would have to maintain the position that the person living in an experience machine lead a good life.

Notice that the argument isn't saying that experiences don't matter to welfare or the good life. Rather, the argument is saying that the character of our experiences is not all there is to a good life. That is, good experiences might be necessary to a good life, but they are not sufficient. What the experience machine argument is supposed to show is that one of the elements that's missing in an experience-based approach to welfare is the notion of experiences being *real*.

Even if you don't find the experience machine argument compelling, there are other issues that remain. My experience of listening to heavy metal may produce pleasure, while it may produce pain in you. The fact that it feels good to me but bad to you suggests that we can't be having the same experience. If that's right, then we have to say that there are two different types of experiences being had while you and I listen to the same heavy metal track.

If that's not right, that is, if there is only one type of experience being had, is there anything that can be said to explain the difference in how you and I perceive the goodness and badness? One attempt

might be to distinguish between lower and higher order experiences. That is, we might be having the same lower order experience of the music (this might be something like the 'raw' perception of the music) but we could be having different higher order experiences. I might have a higher order experience of the lower order experience of violent drumming as an expression of existential angst, which feels good to release, while your higher order experience of the drumming might remind you of the frustrations of learning to ride a bike.

There are some good reasons for thinking that our experiences have these multiple levels happening in parallel. The smoker might know very well that they are engaging in risky behaviour, but at the same time that risky behaviour could be providing them with a higher order thrill. That might also be what's going on for some people when they watch horror movies.

But not all experiences seem to have such multiple levels. When I taste a scotch whisky from Islay (Scotland), I think the the salty, smoky, burnt rubber flavours are *good*, while many people think those exact same flavours are *bad*. And this difference doesn't seem to be one characterized by differences in higher order experiences. That is, our differences in tastes of scotch seem to be just at the lower order. If it's right, then the challenge for the experienced based view is to resolve the contradiction that the experience is good for one person but bad for another (given that the bad and the good are identified with experiences).

There are additional concerns with the experience-based view concerning the nature of time. A good experience today seems better than the same good experience 10 years in the future. This is not unlike the case of how money works, but it is more general and called *discounting the future*.

Suffice it to say here that a purely experience-based view doesn't distinguish between the goodness or badness of an experience relative to your current point in time, and yet we do seem to behave in such a way that the present and near future are somehow different than the distant past or future.

4.2.3 *Objective List Theories*

Objective list theories start with the recognition that welfare may not be an aggregation of something simple, like a collection of pleasurable experiences. Rather, objective list theories recognize that welfare is a heterogeneous collection of things. We might have something like the following list:

- Good health

- Shelter
- Being in loving relationships with family and friends
- Experiencing beauty
- Engaging in virtuous behavior
- Being rational
- Acquiring knowledge
- Doing things that make a life good, including things on this list

The approach to welfare as a multi-faceted concept is supposedly a strength of objective list theories, but it can also be seen as its primary weakness. There are a few ways to making the objection.

In order for any of the decision making rules to work, they need to assign a number to an outcome. So somehow, the information that's on a list has to be converted into a single unit of measurement - a utile. A list by itself doesn't tell us how to aggregate that information. So in order for an objective list theory to work for the decision making rules we've considered, it has to also provide such a procedure.

THERE ARE REASONS for thinking that there is no plausible procedure for converting lists into utilities. For one, the very items on a list don't seem to be fixed. Different people might have different items on the list (perhaps we should then call it *subjective* list theory). Or, even if the items on the list are the same, different people might weight the items differently. An academic and an athlete are likely to give different weights to acquiring knowledge and good health. So a procedure for converting lists has to deal with the fact that lists might differ in both the items that are on them and the weights to assign to items.

In order for any procedure to have the capacities just stated, there has to be a way for items on the list to be compared. The main thrust of the objection capitalizes on this requirement and it can be represented in the form of a dilemma.

ON THE FIRST HORN, there is no way of making comparisons between the items such that they can be aggregated into a single number. On this horn objective list theories concede that they can't provide an account of utilities for the purposes of decision making.

ON THE SECOND HORN, there is a way of making a comparison between the items. Comparisons are only possible if the items have something in common, some way of relating them to one another. But if they have something in common, then why not use whatever that is as a basis for utilities? For example, if the thing in common between experiencing beauty and being in a loving relationship is

that they produce pleasurable experiences, then why not just use the experienced-based approach in the first place? On this horn of the dilemma, objective list theories give up the very idea that motivated them in the first place: that welfare is heterogeneous and multi-faceted.

IN BRIEF: if welfare is so heterogeneous that the items on the list cannot be compared, then the first horn is faced, but if items can be compared, then the second horn is faced. On either horn, objective list theories aren't the right account to provide a measure of utility.

4.2.4   *Preference-based Theories*

ONE OBJECTION COMMON to both objective list theories and experience-based theories is that different people can have different views on what is good for them. That is, what counts as good for me can differ from what counts as good for you. One way to think of this is to say that welfare is something that is subjective, not objective.

Preference-based theories take this kind of objection seriously. On these accounts, what counts as good is that a preference is satisfied. *A* is better than *B* for an agent if and only if *A* is preferred to *B* by that agent. If we are talking about some other agent where *B* is preferred to *A*, then for that agent *B* is better than *A*.

Before we consider some complications of this approach, it's worth noting several advantages. As we already noted, a preference-based theory can handle the issues we raised earlier: that what increases welfare for one person (like our athlete) may not be what increases welfare for another (like our acaedmic).

But another substantial advantage is that preferences can be both nuanced on the one hand, and provide a unifying measure that we can use for utilities. For example, it has no problem handling differences in how people weight different items, or the times at which they would like for the items to be instantiated. For example, one person may prefer a steady stream of somewhat pleasurable experiences, while another enjoys a roller coaster of ups and downs in life. Preference-based theories are sufficiently general to account for such differences.

One serious complication of this view, however, comes from the idea that people could be wrong about what's good for them. On a purely subjectivist view, whatever a person desires is what's good for them, and vice versa. It seems perfectly reasonable, however, to say that sometimes people don't know what's good for them. Sometimes we think we know what's good for us, but we're actually

wrong about that. Sometimes our preferences are challenged by other people. Anyone that has had the thoughts, "I wish I didn't drink so much" or "I wish I didn't smoke" are good examples of this.

So the trouble is, we want to be able to say both how welfare can be differential across individuals, but at the same time we want to be able to account for the fact that individuals could be wrong about what's good for them. Surely the right answer is somewhere in between. We don't expect utilities to be entirely subjective, nor entirely objective. Call this the subjective-objective *tension* (as opposed to dilemma, which would suggest we have to pick one or the other).

ONE OF THE ADVANTAGES of the preference-based approach is that it needn't be purely subjectivist. In fact, one of its chief advantages is that we needn't wait for a resolution in the debate about how to balance the two sides of the subjective-objective tension. All that a preference-based approach demands is that some *minimal* criteria be met about what preferences are *for an individual*. Decision theory, as we're considering it, is about how an individual should decide *given their preferences* (recall our discussion of instrumental rationality). It thereby sidesteps some of the issues regarding interpersonal comparisons. In the previous chapter we looked at two minimal criteria for preferences (transitivity and completeness) and arguments for them (money pumps).

4.3   *Applications and Challenges*

Whatever view you might be attracted to about cardinal utilities, let's have look at an application of them that is not available if we just used ordinal utilities. This is called the *minimax regret rule*. After examining that rule, we consider some challenges of trying to distill down utilities into a single number.

4.3.1   *Minimize Regret*

A simple thing we can do with cardinal utilities is give a rough measure of regret. Let's consider an example. Suppose your options are to party, study, or call your family. Suppose further that tomorrow's exam will be either very difficult, medium difficult, or easy. Let's say that you assign the following utilities to the nine possible outcomes.

	Very Difficult	Medium	Easy Exam
Party!	1	5	20
Study!	15	10	3

	Very Difficult	Medium	Easy Exam
Call Family	7	8	9

There are different ways of thinking about regret. Here's one plausible way. Given a particular state, the amount of regret of an option is the difference between the highest outcome among all the options (under the given state) and the option under consideration. For example, under the state Easy Exam, the option to party has the highest outcome. So, the option to party has a regret of 0, the option to study has a regret of 17, and the option to call family has a regret of 11. Using the same strategy, we can calculate the amount of regret for the options under each of the other states. We then get the following **table of regrets**:

	Very Difficult	Medium	Easy Exam
Party!	14	5	0
Study!	0	0	17
Call Family	8	2	11

Using this regret table, the **minimax regret** rule says to *min*imize the *max*imum possible regret. The strategy proceeds in two steps: first, find the maximum regret for each option (for Party! it's 14, for Study it's 17, and for Call Family it's 11), then second, pick the option with the lowest regret (in our example it's Call Family).

The minimax regret rule seems intuitive enough. Given the preference assignment, we really prefer to party if the exam is easy, and we really prefer to study if the exam turns out to be very difficult. The option to call family lies somewhere in the middle: no matter how hard the exam turns out, it feels pretty good to get encouragement from talking with family. But we are uncertain about how difficult the exam will be. Because we're uncertain, and under the assumption that we would rather not regret our choice, our best bet is to call family.

Independence of Irrelevant Alternatives

There is a principle called *independence of irrelevant alternatives*, or IIA. The basic idea is that if you choose a specific option, like Call Family above, then removing one of the other options shouldn't change the recommendation. More formally, if option x is recommended among a set of options T, then if S is a subset of options from T and x is a member of S, then x should still be the recommended option.

It turns out that the minimax regret rule does not obey IIA. If it did, then that means the option to call family would still be recom-

mended even if we ignored the option to go to the party (maybe it was canceled last minute). But we can show that minimax regret actually changes its recommendation.

To show this, we first take a subset of our previous options and make sure that Call Family is still a member of that subset. We can do this by deleting the option to party from the table above. That gets us this next table.

	Very Difficult	Medium	Easy Exam
Study!	15	10	3
Call Family	7	8	9

Next we create our new regret table. Notice that this time we'll get different numbers under the Easy Exam column. This is because the highest outcome is no longer the one where we party and the exam is easy (20) - we've removed the party option altogether and thereby the corresponding outcomes. So now the highest outcome under Easy Exam is 9. In the other two columns, the highest outcomes are still the same (they are the study row). When we create the regret table, we then get the following.

	Very Difficult	Medium	Easy Exam
Study!	0	0	6
Call Family	8	2	0

Following the minimax regret rule, we see that Study has a maximum regret of 6 while calling family has a maximum regret of 8. Picking the minimum of these leads to the recommendation of studying, the opposite of the recommendation it gave when we considered that the party was still on.

If we think that the Independence of Irrelevant Alternatives principle is correct, then that speaks unfavourably to the minimax regret rule. It is not obvious, however, that we should accept IIA. Here are two arguments why we might not accept IIA.

THE FIRST ARGUMENT AGAINST IIA comes from empirical considerations. People seem to systematically behave at odds with IIA. A famous example is known as *the decoy effect*.[10] Suppose you are deciding between two sandwiches, a small one (say a 6 inch loaf) vs a large (say 12 inch). The small is $6 and the large is $10. Between these two options, let's suppose that your preference is for the small sandwich - it's enough to fill you up and costs less than the large sandwich. Now imagine that the store announces a limited time

[10] CITE Kahnemann, CITE Ariely

medium sandwich, coming in at 9 inches and $9. A lot of empirical work suggests that your disposition towards the large sandwich changes with the introduction of the medium sandwich, even if you don't actually have a preference for the medium sandwich. That is, in many cases people will opt for the large sandwich when the medium sandwich is present, but they prefer the small sandwich when it is not, even though the medium sandwich is not preferred. The medium sandwich is called the decoy.

Once you see it, you'll start noticing the decoy effect in a lot of places, especially in advertising. The effect is not restricted to marketing, however. Its applications are broad, even in dating. Suppose you see Jordan and Taylor at a party and you have a romantic interest in each of them. You find both equally attractive. You know Jordan is a great piano player while Taylor is a great guitar player. You have a bit of a soft spot for piano players, so you're inclined to strike up a conversation with Jordan (i.e., you have slight preference for Jordan over Taylor). But just as you head in Jordan's direction, you hear Taylor strumming along to a tune with one of their guitar playing friends, Kennedy. It's clear that Taylor is a more talented player than Kennedy. Research suggests that because you can compare Taylor to Kennedy - Kennedy is the decoy in this case - your preference can switch from your interest in Jordan to Taylor. Kennedy is irrelevant - you have no romantic interest in Kennedy - but their mere presence can affect your preferences. Of course, you hope that you're not the decoy for the friend who invited you to the party.[11]

[11] CITE Dan Ariely

Just because the decoy effect is widespread, however, does not make it a definitive point against IIA. It is possible to say, in IIA's defense, that people are systematically failing to be rational. How much one takes seriously the empirical data in informing theoretical choices will depend in part on how descriptive and normative aspects of decision theory interact. As a descriptive theory, IIA fails to hold, but as a normative theory we may endorse IIA and point to the decoy effect as a systematic failure on our part to be ideal decision makers (just like some of us systematically fail at parallel parking).

The second argument against IIA, however, comes from theoretical considerations. The thought is that even as part of a normative theory, IIA is too strict of a requirement for rationality. One might think, for example, that when we are considering our options, we should do so holistically. That is, we should not just be comparing options pairwise or side by side until we have a winner. Rather, we should be comparing options as an entire collective. Put differently, when we are considering an option, we should be comparing it to the entire set of alternatives. If that's right, then even the ideal decision

maker would violate IIA.

For example, suppose you collect vinyl records and you're on vacation somewhere to visit a vinyl shop. You see two records, both of which tend to be difficult to find. Let's say you prefer record A over record B. As you head to the counter with A in hand, you overhear some people talking about a stash of counterfeits that have been made of A. Such a copy, call it C, is widely available, is also at the store as a kind of gimmick, and it's noticeably "cheaper looking" than A. So let's say you prefer A over a copy C, and you also prefer record B to C. To many people it seems perfectly reasonable to think that the popularity of C makes A look less attractive, such that your preferences shift to prefer B over A. After all, B is more unique, seeing as it's not the sort of thing that there are as many copies of. (Similarly, we might run the story the other way so that instead of there being copies of A, there are lots of copies of B, but this time the fact that there are more copies makes B more special to you because you would have an original.)

We'll see other theoretical arguments against IIA in more detail when we look at the Allais paradox. To appreciate that argument, we first need to understand the rule for maximizing expected utility, which we'll cover in the next chapter. The key difference we'll see between maximizing expected utility and the rules we've looked at in this chapter has to do with how they incorporate uncertainty. But first, let's look at one more application of cardinal utilities.

4.3.2   *Multiattribute Approach*

In a single-attribute approach, we are forced to make direct comparisons when making decisions. This leads to us making comparisons between things like lives and money: at what point is the dollar cost of a search and rescue too high to save the lives of a few teenagers lost at sea? Here we are placing money and lives and a single utility scale.

In a multi-attribute approach, we allow for there to be multiple scales. Which scales are used depends on the attributes relevant to the decision. In the above example, we might use money to measure the financial costs of a search and rescue, and use number of lives saved as the scale for human welfare. The next step is then to determine how to aggregate the attributes into some overall ranking.

Here's an example. Suppose a politician is deciding on extending the search and rescue mission for the teenagers lost at sea. There are three attributes under consideration: the cost in dollars for each day of search, the number of lives saved, and the political implications (e.g., getting or losing votes). The politician is considering three op-

tions: end search, extend search one week, extend search indefinitely. For simplicity's sake, we're going to imagine that there's only one overarching world state (say, that the teenagers are still alive). For each outcome, the politician could use an ordinal ranking (i.e., a 1 means least preferred, not 1 life saved). Note that even though there is only one world state at the moment (Teens Alive), the outcome for End Search and Teens Alive has three compartments, one for each of the three attributes.

	Money	Lives	Votes
End Search	3	1	1
Extend One Week	2	3	2
Extend Indefinitely	1	2	3

Unfortunately for the politician, there is no dominant option; dominance reasoning will not provide guidance for what to choose. The politician is not limited to the ordinal scale, however. We could instead use an interval (or quantitative) scale. By interpreting the numbers in the above table as cardinal numbers, we can then weight the values of the outcomes with how important each attribute is.

Weighting by additivity

In order to assign weights, we need to assume that they are additive and that the sum of the weights is equal to 1. Let's suppose that each attribute is equally important as any other. Then since we have three attributes, each one is assigned $1/3$. The value of an option for the overarching outcome is then the weighted sum of its subdivided outcomes. For example, the option Extend One Week has a value of $2 \times 1/3 + 3 \times 1/3 + 2 \times 1/3 = 2.333$. Performing the analogous calculations for the other two options reveals that Extend One Week has the maximum value.

	Money ($1/3$)	Lives ($1/3$)	Votes ($1/3$)	Aggregate
End Search	$3 \times \frac{1}{3}$	$1 \times \frac{1}{3}$	$1 \times \frac{1}{3}$	$=1.667$
Extend One Week	$2 \times \frac{1}{3}$	$3 \times \frac{1}{3}$	$2 \times \frac{1}{3}$	$=2.333$
Extend Indefinitely	$1 \times \frac{1}{3}$	$2 \times \frac{1}{3}$	$3 \times \frac{1}{3}$	$=2.000$

Suppose instead that Votes is the most important thing for the politician. We might say, for example, that votes are twice as important as lives, and lives in turn are twice as important as money, i.e., Votes are weighted $4/7$, Lives at $2/7$, and Money at $1/7$. Using these weights the Extend One Week option has a (rounded) value of 2.286, but the Extend Indefinitely option has a (rounded) value of 2.429. So different weightings of the attributes can lead to different recommendations.

	Money (1/7)	Lives (2/7)	Votes (4/7)	Aggregate
End Search	$3 \times \frac{1}{7}$	$1 \times \frac{2}{7}$	$1 \times \frac{4}{7}$	=1.286
Extend One Week	$2 \times \frac{1}{7}$	$3 \times \frac{2}{7}$	$2 \times \frac{4}{7}$	=2.286
Extend Indefinitely	$1 \times \frac{1}{7}$	$2 \times \frac{2}{7}$	$3 \times \frac{4}{7}$	=2.429

Don't forget that the columns in this table are all under the state assumption of Teens Alive. You'll have something similar for the state Not Alive. Later we'll see how we can combine states and utilities by using probabilities, but that's in the next chapter.

Non-additive strategy and general challenges

This weighting strategy of the multi-attribute approach assumes that the attributes are additive, i.e., that their relative importance can be summed. Consider again a criticism of the completeness axiom, which can be launched similarly here. The aggregation of attributes on a scale from 0 to 1 that sum in total to 1 is no different than comparing objects on a single scale. It's the same problem, we've just moved from comparison of objects to comparisons of attributes.

Using non-additive criteria is one strategy for sidestepping the above objection while still handling incomplete preference orderings. Suppose each attribute has an *aspirational level* such that if that minimum level is not achieved, then the corresponding option or alternative is disregarded, even if that option ranks highly in other attributes. For example, suppose that we set the aspirational level for each of the attributes above at 2. Then End Search and Extend Indefinitely will be disregarded because each of them has at least one outcome below 2. Notice that on this strategy the attributes aren't being compared on some single scale. Each attribute gets to set its own level and then we check to make sure that the options meet that minimum level.

A severe complication of this non-additive strategy is specifying aspirational levels for each attribute: there seems to be both arbitrariness and vagueness to the levels.

There are concerns related to multi-attribute approaches generally. Without some theory for individuating attributes, it is possible that the strategy leads to conflicting recommendations (whether or not they are additive). For example, suppose that a medical products agency is deciding whether to approve a new hand cream. It is unknown whether the long term effects of the hand cream leads to increased toxicity. Toxicity may then be taken as a relevant attribute. Let us suppose that while for half the population the hand cream increases toxicity, for the other half of the population the cream actually leads to a decrease in toxicity. We can suppose that whether it does or doesn't depends on something genetic. If we conceive of

toxicity as a single attribute, then the average increase in toxicity is zero because one half of the population cancels out the other. By these lights, the hand cream might get approved. It seems reasonable, however, to characterize toxicity in terms of two attributes: "toxicity for people with gene A" and "toxicity for people with gene B". If we individuate attributes in this way, the recommendation may be to not approve the hand cream if the risk for people with one of the genes is sufficiently large. So here is an example where there are two different ways of aggregating risks that lead to conflicting recommendations, and since we have no way of determining which is the correct way of individuating attributes, a multi-attribute approach does not solve the problem of dealing with incomplete preference orderings.

4.4   *More Challenges and Final Remarks*

There is much more to be said about the notion of utilities. Even in the more minimal case of preference-based theories, we assume that they are transitive and complete. Moreover, we have implicitly assumed that a person knows what their own preferences are, at least in the moment. But there are good reasons for thinking that in many scenarios we *don't* know our own preferences and look towards others to either discover or create our own. We also have reasons for thinking that preferences can change over time, and in some cases it seems that the way our preferences change is not predictable beforehand.

Even if individuals had unchanging preferences that they knew perfectly, there are good reasons for treating utilities across time differently. It seems plausible to think that a unit of utility is more valuable today (from your perspective on the present day) than it is ten years from now. After all, there's no guarantee that you'll live another ten years to obtain that unit of utility, and it's far more likely that you'll live through today. Investing money illustrates a similar idea quite nicely. $1,000 dollars today is worth more than $1,000 next year (even if we increase this amount to account for inflation) because you can invest it today and turn it into more than $1,000 by next year (market crashes notwithstanding). For these sorts of considerations, we often *discount* our future selves in the decisions we are making in the present. That is, while it is important to think about yourself in the future, the preferences of your future self should not be given the same weight as you in the present. Again, who knows if your future self even comes to be: as the famous economist Maynard Keynes said, "In the end we are all dead."

Interesting, it's not clear whether we should apply the same sort of thinking when we are considering decisions at the aggregate or

group level. Unlike the guarantee that your life will come to an end and we have a decent sense of how long you will *not* live, the life of communities is far more open-ended. In fact, society is likely to go on far longer than individuals that it makes sense to model society as if it were to go on forever - this is sometimes referred to as the 'infinite life' assumption in economics. Without an endpoint in sight, it's not clear how much we should be discounting future generations. In fact, some philosophers and economists suggest that, given that we have some moral obligations to future society, we shouldn't discount future generations at all. Frank Ramsey said, "In time the world will cool and everything will die, but that is a long time off still, and its present value at compound discount is almost nothing." So while the death of individuals warrants the application of a discount function, the death of society is so far off that the "same'' reason does not apply here. Once again we see there is a kind of discontinuity between decision making at the level of individuals and the level of groups.

All of these considerations raise interesting questions about how preferences affect decision making. But we need to return to the idea of an *expectation* in the first place. To do that, it is not enough to just think in terms of utilities. We have to be more explicit about the notion of a probability, like we used in lottery option procedure, and how they are used in models of decision making. We turn to that next.

Exercises

1. Suppose we have a utility scale that goes from 0 to 5 and we're trying to determine what utility Alice assigns to an option called *D*. We know that Alice has the preference ordering of $A \succ D \succ C$. Suppose we present Alice with a lottery option *L* and we determine that Alice said $D \sim L$ when *L* consists of 60 tickets of *A* and 40 tickets of *C*.

   • What utility does *D* have for Alice assuming that *A* is 5 and *C* is 0?
   • What if instead *A* is 4 and *C* is 3?

2. Which interpretation of cardinal utilities (hedonism, objective list theory, or preference-based) do you find most compelling? Why?

3. Provide an application of the minimax regret rule to one of your own decisions. Remember you will need two tables, one is the standard utility table, the second is the table of regrets (which you calculate from the utility table).

- Does the recommendation of the minimax regret rule correspond with your intuition about what to do?
- Does your example obey the principle of independence of irrelevant alternatives?

4. Provide an application of the multiattribute approach to one of your own decisions, using weighting by additivity.

- Does the recommendation of the multiattribute approach correspond with your intuition about what to do?
- If not, is there any way to update the weights so that it would?
- If so, how much could you change the weights so that the approach would make a different recommendation?

5
Expected Utilities

We developed the lottery option procedure to help us build interval scales for people's preferences, which lets us both order preferences and say how much more one thing is preferred to another.[1] But we noted a curiousity about this procedure. Saying that you prefer a lottery is quite a different thing than saying you prefer a particular outcome, like a book. It seems intuitive enough to think about selecting book B rather than another: we purchase it, put it beside our bed, read it, etc. But what does it mean to act on or choose a lottery? Yes, you can purchase a ticket. But you can't *choose* the *winning* ticket in the sense that your choice guarantees the ticket is a winning one. Whether your ticket wins or loses is the outcome of some chancy process. How can we make better sense of this?

The central concept that will help us here is *expectation*. Informally, we think of expectations as strong beliefs that something will occur. When walking I almost always expect that the ground will support my weight. When I set my alarm I expect it to wake me up. When I work at my job I expect to get paid. These examples are in contrast to things we might *merely* hope or desire to be true.

There is also a more technical meaning of what it is to have an expectation. Here we make use of the concept of probabilities in making a kind of measured prediction. We move away from a strong belief that something *will* occur, and more towards what could occur to some degree of chance.

5.1   *Expected Utility by Example*

Here is an example. Suppose you have a rash on your hand. In half of all cases the rash goes away by itself, for the other half it doesn't. So, if you do nothing, there's a 50/50 chance the rash gets better. There's an over the counter cream you can use that helps the rash get better, but it's not perfect: 85% of the time the rash goes away, but there's a 15% chance that it continues even when using the cream.[2]

[1] Recall how the lottery option procedure worked: given options A≻B≻C, where A and C make up a lottery L, what percent of L's tickets need to be A in order for the person to be indifferent between A and B? The percent of A tickets is the percent of B's desirability of A.

[2] Notice that we're not being as careful like we have in the past. Here the "rash gets better" state is at least *somewhat* in our control (our choice affects the probabilities of the outcome). But there's also aspects of the cream that are outside of our control, like the sensitivity of your skin and the efficacy of the cream given your skin. It's arguably a feature that we don't have to be as careful, but we'll return to this

	Rash gets better	Rash continues
Use Cream	85%	15%
Do Nothing	50%	50%

The notion of expectation that we are going to use combines this kind of information about probabilities with cardinal utilities. In fact, we'll have a new strategy called **maximize expected utility**, which will use information about how probable outcomes are, like in the table above, in combination with our utilities of those outcomes. That is, we use the probabilities *to weight* a preference assignment.

To make the example easier, let's think of utilities in terms of money.[3] There are two costs we need to think about: the cost of the over the counter hand cream and the cost of seeing the doctor if the rash continues. (Strictly speaking, there is also a third "cost" if we assume that the price of doing neither is $0.) Let's say that the cost of the hand cream is $20 and the cost of going to the doctor is $100. In the event that you tried the hand cream and it didn't work, you still have to go to the doctor (who may prescribe a stronger cream), so you're paying both costs. In terms of a preference assignment, we then get the following table.

[3] Some authors will distinguish between *expected monetary value* and *expected utility* in order to acknowledge the concerns we raised in the previous chapter. Here we're just trying to highlight how to think about *expectation*, and so for most of this chapter we'll treat utilities and monetary value the same.

	Rash gets better	Rash continues
Use Cream	-$20	-$20+(-$100) = -$120
Do Nothing	-$0	-$100

If all that matters are these utilities, then we can apply the rules we used in the previous chapters. For example, dominance reasoning would tell us not to use hand cream because for each state, the option not to use cream is better (i.e. less expensive) than the option to use cream.

But utilities are not all that matter. We also have other information: the probabilities of each of these outcomes.

The rule for maximizing expected utility proceeds in two steps:

- Step One: combine the utility and probability tables to calculate the expected utility of each option.
- Step Two: select the option with the maximum expected utility.

Let's go through these separately.

5.1.1 Step One: Compute EU of each option

Here's how we calculate the expected utility of an option. First, for each outcome, multiple the utility of that outcome (using the utilities

table) by the probability of that outcome (using the probability table).
Second, take the sum of these products.

For example, to calculate the expected utility of using cream, I
would first do this:

	Rash gets better	Rash continues
Use Cream	-$20 x 0.85	-$120 x0.15

Which gives me this:

	Rash gets better	Rash continues
Use Cream	-$20 x 0.85 = -$17	-$120 x0.15 = -$18

I then take the sum of these outcomes, which is -$35. We say that
the expected utility of using the hand cream is a *loss* of 35 dollars
(recall we're ignoring a bunch of complications for now, like whether
the hand cream makes my hands feel greasy, how my hands smell,
etc.). We'll represent this expected utility as **EU** and put it in as col-
umn to the very right, being careful not to interpret that column as
another state.

	Rash gets better	Rash continues	EU
Use Cream	-$17	-$18	(-$17) + (-$18) = **-$35**

We then do the same computation to calculate the expected utility
of Do Nothing. So our complete table would look like this (showing
the work for Do Nothing):

	Rash gets better	Rash continues	EU
Use Cream	-$17	-$18	**-$35**
Do Nothing	-$0 x 0.5 = $0	-$100 x 0.5 = -$50	(-$0)+(-$50) = **-$50**

This makes up our first step.

5.1.2   *Step Two: Find maximum EU*

Once we have our table where we have calculated the expected util-
ity for each option, we then look for which one is the maximum. We
have to be a bit careful here, because the signs of the numbers mat-
ter, not just the magnitude. So while $50 is more than $35, what we
are concerned with is avoiding the loss of money. In other words,
when utilities are positive we want to take the largest one, but when

utilities are negative we want to take the smallest one. In the current context the utilities are negative, so we are trying to maximize the least amount of loss. Hence the option with the maximum expected utility is Use Cream. We'll sometimes put an explanation point beside the maximum expected utility number.

	Rash gets better	Rash continues	EU
Use Cream	-$17	-$18	**-$35!**
Do Nothing	$0	-$50	**-$50**

If two options have the same expected utility, and all other expected utilities are smaller than those options, then there is a tie. A common suggestion is that such ties mean that either option is equally as good (or least bad) as the other. We will see, however, that we might want to be more nuanced than merely maximizing expected utilities. For example, it's possible that two options have the same expected utilities, but one of them has a higher "risk" in that one of the outcomes has a very low utility, but also a very low probability. We will consider such possibilities shortly when we examine criticisms of the strategy of maximizing expected utilities.

5.2   (MEU) Maximize Expected Utility Strategy

It will be helpful to have some notation. Let's say that $U(A|S_1)$ means "the utility of option A given state 1''. So we could write U(Use cream|rash gets better) to refer to the $20 cost. Or we could write it like an equation. Here's how we might refer to the case where we use cream and the rash continues:

$$U(\text{Use cream}|\text{rash continues}) = -\$120$$

In addition to using U to refer to utilities, we also need a handy way to refer to the probabilities of outcomes in a probability table. Let's use $P(A|S_1)$ to mean the probability of the outcome when option A is selected and state 1 occurs. So in order to refer to the chance that the rash gets better given that you use hand cream, which is 85%, we would write

$$P(\text{Use cream}|\text{rash gets better}) = 0.85$$

Now that we have a way to refer to utilities and a way to refer to probabilities, we can write down a general way of how expected utilities are calculation. The expected utility of an option (say A) where there are two states can be written like this:[4]

[4] Don't forget your order of operations!

$$EU(A) = P(A|S_1) \times U(A|S_1) + P(A|S_2) \times U(A|S_2)$$

Plugging in the numbers from the table above for the hand cream option, we would get this:

$$EU(\text{Use Cream}) = (0.85 \times -\$20) + (0.15 \times -\$120)$$
$$= -\$35$$

Similarly, the expected utility for not using the hand cream would be:

$$EU(\text{Do Nothing}) = 0.5 \times -\$0 + 0.5 \times -\$100$$
$$= -\$50$$

As illustrated in the above example, once we have calculated the expected utility for each option (i.e. each row in our tables), then we pick the maximum of all the utilities. Note that this strategy works whether there are just two options, or if there are three, four, etc. The first step is a calculation of the expected utility for each option, which is done independently for each row. So the addition of another row doesn't make a difference to how the other options were calculated (well almost - we'll see later that we're making an important assumption here!). It's only in the second step that might be affected, but here we need to have calculated the expected utility for each option before we can go about selecting the maximum one.

Note that we have a difference in recommendations from some other rules. Dominance reasoning would tell us not to use the hand cream. But dominance reasoning ignores some important information about the chances of the different outcomes happening. One way to incorporate that information, in fact the most common way to do so, is to weight the utility of the outcome with the probability of that outcome.[5] When we do that for each outcome for a given option and then add them up, we get the expected utility for that option. If we then compare the expected utilities of each option, at least for the example we've considered above, we get a different recommendation: don't use the hand cream.

The *maximize expected utility* rule is the predominant strategy in the field of decision theory.[6] It is a very rich strategy. For example, we can actually recover the recommendation that dominance reasoning gave us. In order to do that, we ignore the information that using hand cream improves the chances of the rash getting better. We can represent this ignorance by giving each outcome for a given option the same probability. That is, we use the following table of probabilities:

[5] The idea of "weighting" is what the operation of multiplication is doing here.

[6] That's not to suggest it's always as simple as we've been illustrating, but it is at the heart of many decision analyses.

	Rash gets better	Rash continues
Use Cream	0.5	0.5
Do Nothing	0.5	0.5

Notice that the probability in a given row add up to 1. That's because the states are *exhaustive* in each given option. That is, if I use the hand cream, my rash will either get better or it won't - there are no additional things that can happen. The same is true if I don't use it: either the rash goes away or it doesn't.

Now if we plug these numbers into our expected utilities, we get

$$EU(\text{use cream}) = 0.5 \times -\$20 + 0.5 \times -\$120$$
$$= -\$70$$

and

$$EU(\text{don't use cream}) = 0.5 \times -\$0 + 0.5 \times -\$100$$
$$= -\$50$$

Notice that when we take the maximum of these[7] then the *maximize expected utility* rule says not to use the hand cream. That's precisely the recommendation that dominance reasoning originally gave us!

The main takeaway is this. When we are in the context of making a decision under ignorance where we don't know what the probabilities of outcomes are, the dominance principle is usually the "go to" rule that is first tried (and sometimes it doesn't give us a recommendation). In the context of making a decision under known risk, we have information about the probabilities of outcomes and we use that information to calculate expected utilities, which in turn are maximized. Interestingly, we can use the decision rule of maximize expected utility to also represent the dominance strategy. We do this by "ignoring'' the probabilities of outcomes by making them all equally likely. In a rough way, then, maximize expected utility (MEU) is a generalization of the dominance principle, at least in terms of the recommendations MEU makes.

We say "in a rough way" because the recommendations that the dominance strategy and MEU make can come apart. In fact we have already seen this in the introduction. Recall *Newcomb's Problem* where you can either pick the contents of just box A or the contents of both box A and box B. The utility table looked like this:

[7] Remember again these are representing *costs*.

	AI predicts two box	AI predicts one box
One box (just A)	$0	$1,000,000
Two box (A and B)	$1,000	$1,001,000

Dominance reasoning makes the recommendation to select the Two Box option, because whatever the AI predicted, you get an extra $1,000. But dominance reasoning ignores the probabilities of each of the outcomes. In the description of the setup we said that there is a very strong correlation between what options people pick and what the AI predicts. So if you pick the One Box option, then the probability that the AI predicted that you would pick that option should be high, very high. It might not be perfect, but we can suppose that the AI is right 99% of the time. So, the probability table might look like this:

	AI predicts two box	AI predicts one box
One box (just A)	0.01	0.99
Two box (A and B)	0.99	0.01

Following the steps we outlined above for the MEU stragey, we get the following expected utilities for each option:

$$EU(\text{OneBox}) = (0.01 \times \$0) + (0.99 \times \$1,000,000)$$
$$= \$990,000$$

$$EU(\text{TwoBox}) = (0.99 \times \$1,000) + (0.01 \times \$1,001,000)$$
$$= \$990 + \$10,010$$
$$= \$11,000$$

Let's add these to our utility table.

	AI predicts two box	AI predicts one box	EU
One box (just A)	$0	$1,000,000	**$990,000!**
Two box (A and B)	$1,000	$1,001,000	**$11,000**

Clearly the One Box option has a much higher expected utility than the Two Box option! So, in contrast to the dominance strategy recommendation, the MEU strategy recommends that you should select just box A.

We have in a sense made interesting progress, and in another sense we have not. We have made progress in the sense that we can be more precise in how we describe the two different ways of reasoning about what do to in the Newcomb Problem. We have not yet

made progress in providing an argument for whether we should use dominance reasoning or MEU. But the fact that we can be more precise about the two reasoning processes means we have a better understanding of what background assumptions each one is making. In turn, by analyzing those assumptions we can consider arguments for and against each of the two strategies. Much of the progress that has been made in decision theory has come from these kinds of analyses, which we'll cover in due time.

Next, we're going to look at an application of MEU that builds from a previous example.

5.3   Application: Combining MEU and the Multi-Attribute Approach

In the multi-attribute approach we covered, we showed how utilities might come from multiple scales. We used an example of a politician combining three scales (money, lives, and votes) to make a decision about a search and rescue mission for some teenagers lost at sea. The politician is considering three options: end search, extend search one week, extend search indefinitely, and we supposed that there were two possible states: Teens Alive and Teens Not Alive.

It was crucial in that example that we were focusing on just the state Teens Alive. As a reminder, the final utility table for the Teens Alive state was this:

	Money (1/7)		Lives (2/7)		Votes (4/7)	Aggregate
End Search	$3 \times \frac{1}{7}$	+	$1 \times \frac{2}{7}$	+	$1 \times \frac{4}{7}$	=1.286
Extend One Week	$2 \times \frac{1}{7}$	+	$3 \times \frac{2}{7}$	+	$2 \times \frac{4}{7}$	=2.286
Extend Indefinitely	$1 \times \frac{1}{7}$	+	$2 \times \frac{2}{7}$	+	$3 \times \frac{4}{7}$	=2.429

What the aggregate utility in the very right column represents is the "fusion" of all the different scales (money, lives, votes) into a single utility value. To do that we weighted the utility of an outcome by how important the relevant scale (the column) was to the politician. Notice that the weights assigned to the columns here are *not* the same as the probabilities that we use to weight utilities when calculating expected utilities of outcomes. To see this, notice that the decision problem for the politician needs to include both the Teens Alive state as well as the Teens Not Alive state. In other words, the full decision problem is given by the table below, which is only partially filled in using the aggregate values from the table above.

	Teens Alive	Teens Not Alive	EU
End Search	1.286	?	?
Extend One Week	2.286	?	?
Extend Indefinitely	2.429	?	?

In order to fill in the missing utilities under the Teens Not Alive column, we have to first use the multi-attribute approach to find the aggregate utility for each outcome under that state, using the same weighting as we did before. Recall that we supposed the politician thought lives are weighted twice as much as money, and that votes are weighted twice as much as lives. So we know what the weights are. What we need to do is figure out what the utilities are for the three options now that we're considering the state Teens Not Alive. This can be and probably would be different than the utility table we had from before. If the teens are not alive, then it seems that ending the search sooner would be better than later for all three scales: doing a search and rescue costs money, puts additional lives at risk, and voters aren't typically fond of wasting money. With that reasoning, the politician might use the following multi-attribute table for the Teens Not Alive state.

	Money	Lives	Votes
End Search	3	3	3
Extend One Week	2	2	2
Extend Indefinitely	1	1	1

As before, the politician would then weight these to calculate an aggregate:

	Money (1/7)		Lives (2/7)		Votes (4/7)	Aggregate
End Search	$3 \times \frac{1}{7}$	+	$3 \times \frac{2}{7}$	+	$3 \times \frac{4}{7}$	=3.000
Extend One Week	$2 \times \frac{1}{7}$	+	$2 \times \frac{2}{7}$	+	$2 \times \frac{4}{7}$	=2.000
Extend Indefinitely	$1 \times \frac{1}{7}$	+	$1 \times \frac{2}{7}$	+	$1 \times \frac{4}{7}$	=1.000

Now we can use those aggregate utility numbers to fill in the utilities under the Teens Not Alive state in our decision matrix:

	Teens Alive	Teens Not Alive	EU
End Search	1.286	3	?
Extend One Week	2.286	2	?
Extend Indefinitely	2.429	1	?

The maximize expected utility rule will now weight these utilities with the *probabilities* of the states. Here is the place where the politician will need to collect some information about the world. They might consider success rates of past rescue efforts that are similar to this situation. They might consult some survival experts, locals who travel on the water in that area, weather reports, etc. Suppose at the end of all these considerations, the politician's assessment is that Teens Alive is three times more likely than Teens Not Alive, i.e. that the Teens Alive state has a 75% chance while Teens Not Alive has a 25% chance. Now the politician can calculate the expected utilities for each option (rounding to four decimal places):

	Teens Alive		Teens Not Alive	EU
End Search	1.286 x 0.75	+	3 x 0.25	=1.7145
Extend One Week	2.286 x 0.75	+	2 x 0.25	=2.2145
Extend Indefinitely	2.429 x 0.75	+	1 x 0.25	=2.0718

The maximize EU rule then recommends that the politician extend the search for one week.

This example illustrates the richness of the maximize expected utility rule. It helps us organize a variety of considerations about a decision and tells us how to combine these into a recommendation. We first need to determine the utilities of each outcome, which we did using the multi-attribute approach. We then need to assess the probabilities of the states, which we gestured at here and we will return to in more detail shortly. MEU then tells us to weight the utilities by the probabilities of the states to arrive at a final expected utility for each option. There are many ways that real decisions introduce further complications, particularly in how to estimate the probabilities of states. But such complications don't change the underlying principle of MEU. It is no wonder that MEU is one of the most common ways of representing decision problems to make recommendations.

That said, we do have to take some care in the kinds of inputs we give to MEU. In the next section we'll illustrate this using a historical example.

5.4   *Pascal's Wager*

Let's see what happens if we assume that states can have probabilities of one or zero, or what happens if utilities can be infinite. We're going to use Pascal's Wager as an illustration of some concepts, not as an analysis of Pascal's actual argument, which is concerned about the relationship between faith and reason.[8]

[8] The Stanford Encyclopedia of Philosophy has a great entry on Pascal's Wager.

In this wager, the states are either that God exists or God does not exist. The options are that we believe in God or we disbelieve. Let's set up the framework for our decision table:

	God exists (G)	God does not exist (¬G)
Believe (B)		
Don't Believe (¬B)		

So far we haven't said anything about the probabilities of the outcomes nor the utilities. Let's think first about the utilities. Again, we're going to oversimplify for illustrative purposes. If God exists and I am a believer, then I go to heaven and have eternal life. Presumably heaven is a pretty good place where the "party'' literally never ends - it goes on for eternity. So that means that the amount of utility I would get is infinite. Even if there are some bad days in heaven (but that's probably not true) the number of good days will be larger, and so the net amount is an infinite number of positive utilities, i.e., $+\infty$. If, on the other hand, I don't believe in God and God does exist, then I will experience eternal hell, i.e., $-\infty$.

What about the outcomes under the state where God does not exist? In that case my life is finite and the cumulative number of utilities I will experience, positive or negative, will also be finite. Now let's assume that being a believer comes with some costs: a believer volunteers their time when they'd rather be watching sports, or they give away a larger amount of their income to charity than a disbeliever, or they choose not to imbibe when they really would prefer to drink, etc. Let's say for a believer there is a finite net cost $(-c)$ and for the disbeliever there is a finite net gain $(+c)$.

From these assumptions, we have the following utility table:

	God exists (G)	God does not exist (¬G)
Believe (B)	$+\infty$	$-c$
Don't Believe (¬B)	$-\infty$	$+c$

Before we try to add any information about the probabilities of outcomes, notice that the dominance principle does not make a unique recommendation. If God exists, it is better to believe than not to, and if God does not exist it is better not to believe than to believe.

In order to see what maximize expected utility (MEU) recommends, we have to include probabilities for the outcomes so that we calculate expected utilities. As a first approach, let's use the idea

above to make each state equally probable. In that case we would have the following probability table.

	God exists (G)	God does not exist (¬G)
Believe (B)	0.5	0.5
Don't Believe (¬B)	0.5	0.5

If we plug these numbers in to weight the utilities in our preference table, we get the following expected utilities. For Believe the expected utility is

$$EU(B) = P(B|G) \times U(B|G) + P(B|\neg G) \times U(B|\neg G)$$
$$= 0.5 \times (+\infty) + 0.5 \times -c$$

and for Disbelieve it is

$$EU(\neg B) = P(\neg B|G) \times U(\neg B|G) + P(\neg B|\neg G) \times U(\neg B|\neg G)$$
$$= 0.5 \times (-\infty) + 0.5 \times +c$$

We needn't get into too many details about mathematics involving infinity. It's enough to recognize that we can replace the symbol '∞' with an arbitrarily large number. For our purposes, let's say this number is larger than c, even just by one. So let's replace '∞' by 'c+1'. For Believe it would look like this:

$$EU(B) = 0.5 \times (+\infty) + 0.5 \times -c$$
$$= 0.5 \times (c+1) + 0.5 \times -c$$
$$= 0.5c + 0.5 - 0.5c$$
$$= 0.5$$

And for Disbelieve it would look like this:

$$EU(\neg B) = 0.5 \times (-\infty) + 0.5 \times +c$$
$$= 0.5 \times -(c+1) + 0.5 \times +c$$
$$= -0.5c - 0.5 + 0.5c$$
$$= -0.5$$

It may seem that the expected utility for Believe is just a little bit larger than the expected utility for Disbelieve, but remember that we picked a number to replace the '∞' symbol that was only a little bit larger than c. There is no limit to how large this number could be, so in effect the expected utility of Believe is infinitely larger than the expected utility of Disbelieve!

Put differently, the multiplication by some infinite utility completely washes out whatever the finite negative cost is of believing,

and similarly the multiplication of some negative infinite utility will completely wash out the finite positive gain of disbelieving. Hence, by this setup, MEU says we should believe.

Let's consider for a moment an extreme setup where we assign the state that God exists as having a probability of zero. In that case, the left terms for both expected utility calculations are undefined and we are left with a comparison between the finite losses and gains of believing and disbelieving respectively. If we assign the claim of God existing as having a probability of zero, then MEU would recommend not to believe. (Assuming that we can ignore the undefined terms.)

But not so fast. Assigning a probability of zero to a proposition means that we are 100% certain that it is false and we are 100% certain that its negation is true (in this case that God does not exist). While we colloquially say things like "I'm 100% sure that I parked my car in the Scratchy parking lot'' we can't take that as literally true. Strictly speaking, the only propositions we should be 100% sure of are logical tautologies and perhaps simple mathematical claims, i.e., claims that are necessarily true. But the claim that God exists (or not) is not necessary - it is a contingent claim. That is, at least from our eyes, it's *at least possible that* God exists (or does not). And hence, while we might give one or the other a very high (or low) probability, it should never be actually 1 or 0. In short, even the greatest skeptic ought to be giving at least some non-zero value to both states.

With that in mind then, any non-zero probability is going to be still washed out by some infinite utility, and so we are back to the recommendation that we should believe.[9]

There is, admittedly, something awkward or tricky going on here. It seems that the presumption of an infinite utility is doing all the work. None of the other numbers make a difference. As long as we assign ∞ (infinite gain) to the outcome where we believe and God exists, it doesn't matter what else we assign to the other outcomes (with the exception that it's not another positive ∞ of utility). A critic might say that setting $U(B|G) = \infty$ stacks the deck, so to speak. It seems quite possible that heaven, however great, may have a very large positive payoff, but it is not appropriate to say that it is infinite. We may generally insist that whatever utilities we pick, no matter how large, they must always be finite.

If we insist on such a constraint, so that our utilities always have to be some finite number, and in addition insist that our states have to be some non-zero probability, then it is possible that MEU recommends Disbelieve. Here is an example of how that might work.

Suppose that Judy acknowledges that the utility of heaven is quite high. She acknowledges that it could be several orders of magnitude

[9] Again, to be clear, it was not Blaise Pascal's intention to establish that it is necessary that we believe. Some scholars argue that his intention was to show that rationality can neither support faith or a lack thereof.

higher than the value of the cost of believing, but it is not infinite. For the sake of argument, let's say it's 1 million times higher, that is, $U(B|G) = 1M \times c$. And let's say, again for the sake of argument, that the negative utility of hell is the inverse of this, i.e., $U(\neg B|G) = -1M \times c$. Judy's preference table would then look like this:

	God exists (G)	God does not exist (¬G)
Believe (B)	$1M \times c$	-c
Don't Believe (¬B)	$-1M \times c$	+c

If Judy thinks that it's equally likely that God exists as not, then we get the following two expected utilities. For Believe it would look like this:

$$Exp(B) = 0.5 \times 1,000,000c + 0.5 \times -c$$
$$= 500,000c - 0.5c$$
$$= 499,999.5c$$

And for Disbelieve it would look like this:

$$Exp(\neg B) = 0.5 \times -1,000,000c + 0.5 \times +c$$
$$= -500,000c + 0.5c$$
$$= -499,999.5c$$

So again, MEU recommends that Judy believe.

But not all things are equal according to Judy. If we ask Judy what she thinks the probability of God existing is, she may tell us that she doesn't think it is equally likely. Let's suppose she thinks that the probability that God exists is low, very low. She tells us that she thinks there's a 1 in 10 million chance. That is, she thinks $P(G) = 0.0000001$. While this is very small, it is a non-zero probability. Conversely, this means that she thinks there is a very high probability that God does not exist. In fact it would be a 99.99999% chance, or a probability of 0.9999999. Her probability table would then look more like this:

	God exists (G)	God does not exist (¬G)
Believe (B)	0.0000001	0.9999999
Don't Believe (¬B)	0.0000001	0.9999999

Now let's look at what the expected utilities would be. For Believe it is

$$EU(B) = 0.0000001 \times 1,000,000c + 0.9999999 \times -c$$
$$= 0.1c - 0.9999999c$$
$$= -0.8999999c$$

And for Disbelieve it would look like this:

$$EU(\neg B) = 0.0000001 \times -1,000,000c + 0.9999999 \times +c$$
$$= -0.1c + 0.9999999c$$
$$= +0.8999999c$$

In this case MEU now recommends Disbelieve for Judy. How did this happen given that Judy acknowledges that the positive utility of heaven is super high and the negative utility of hell is drastically low? The answer is part of the key insight of the maximize expected utility rule: utilities are to be weighted by probabilities.

What should we say about all this? Does MEU recommend we believe in God or not? Hopefully our walk through this example illustrates that there is no definite answer. It depends. It depends on what we think the utilities are and also what the probabilities are. The maximize expected utility rule requires that we consider both.

Moreover, suppose in Judy's estimation the MEU recommendation is to believe in God, but she just can't seem to get herself to form that belief. For her, the idea that there is some supernatural being might just be too implausible. But maybe, Judy thinks, she can start doing things that might cultivate certain beliefs, like regularly going to sermons and doing daily Bible readings. This possibility highlights a different kind of assumption that we have been making about the timing of making choices and to what extent we are able to voluntarily make them.[10]

For example, we might decide that we want to become a more patient person. That decision, however, is not a one-time event, nor is it the sort of thing that many people can simply choose, like turning on a light switch. Becoming a more patient person may require the development of particular mental habits and learning how to cultivate certain kinds of behaviors. Such decisions are extended over longer periods of time and it is not obvious how MEU (nor dominance for that matter) makes recommendations for such choices.

[10] It also highlights complications about whether beliefs, desires, preferences, and so on, are the sorts of things we can reliable infer from what people claim them to be (their self-reports), or whether we have to look towards their actual behavior (which *always* requires us to make interpretations of them, which themselves can be scrutinized).

5.5 Key Take Aways

Recall that the rules we looked at like dominance reasoning did not pay attention to the probabilities of outcomes. It is a central feature of maximize expected utility that we pay attention to both probabilities and utilities. Doing so means that we have to be particularly mindful. High utilities can be substantially down weighted by low probabilities. And likewise, high probabilities can be down weighted by low utilities. According to MEU it is the *expectation* that should drive the decision making, and the expectation is the sum of utilities *that are*

weighted by probabilities. This is the key insight that moves us from the more simplistic rules to MEU.

That said, while the math and concept underlying MEU is relatively simple, the devil lies in the details of how we estimate the probabilities of outcomes. We have been more or less assuming that we are given the utilities and probabilities and the strategies we've looked at tells us what to do when we have them. But that itself is actually assuming quite a lot. In the coming chapters we will have much to say about the nature of probabilities and how we arrive at them.[11]

Moreover, there are lots of decisions that are more temporally-extended than the sort where MEU is easily applicable. For example, the MEU strategy seems like a useful way of deciding whether to buy a raffle or lottery ticket: we usually know the cost of the ticket, the gain of the prize if we win, and with some effort we can estimate the probability of winning (if we can find out how many tickets there are, for example). But the outcomes of decisions don't always have such clear-cut boundaries, and similarly the places where we "intervene" with our choice isn't a clear localized time point like the purchasing of a ticket. It seems quite natural to ask ourselves questions like, "what kind of person to you want to be or become?" and that choice is a complex sequence of decision making. This can be illustrated starkly in Newcomb's problem. The two-boxer might justify their choice by the fact that the one-boxer is leaving $1,000 on the table. The one-boxer might respond by appealing to a more general character: the type of person that tends to get $1,000,000 is the type of person that picks just the one box, even though that choice in the moment doesn't cause that money to be there.

So there is much conceptual work left for us to do. We'll think about probabilities, assess the range of decision making problems where MEU is applicable, and identify ways in which it might be supplemented. But first, let's examine some arguments for and against MEU more generally.

Exercises

1. Suppose a slot machine pays off $25 a fiftieth of the time and costs a $1 to play, and a video poker machine pays off $10 a twentieth of the time and costs $2 to play. Which machine is the better bet in terms of expected utility (assume here that money and utilities are directly related)?[12]

2. You're considering downloading a new game for your phone. The game costs $0.99. But as a promotion, the first 50,000 downloaders

[11] Admittedly we won't spend much more time talking about utilities and desires - in part because there isn't the same kind of agreed upon theory of them like there is for probabilities and beliefs.

[12] This question is adopted from Weisberg's *Odds and Ends*, Chapter 11, Exercise 3.

are being entered in a fair lottery with a $10,000 cash prize.

If you know you'll be one of the first 50,000 downloaders, what is the expected monetary value of downloading the game?[13]

3. Suppose the government is deciding whether to enact a new tax. If the tax is enacted, it will bring in $700 million in revenue.

But it could also hurt the economy. The chance of harm to the economy is small, just 1/5. But it would cost the country $1,200 million in lost earnings. (The $700 million in revenue would still be gained, partially offsetting this loss.)

Treat gains as positive and losses as negative.

a. What is the expected monetary value of enacting the new tax?

The government also has the option of conducting a study before deciding whether to enact the new tax. If the study's findings are bad news, that means the chance of harm to the economy is actually double what they thought. If its findings are good news, then the chance of harm to the economy is actually half of what they thought.

b. Suppose the government conducts the study and its findings are good news. What will the expected monetary value of enacting the tax be then?
c. Suppose the government conducts the study and its findings are bad news. What will the expected monetary value of enacting the tax be then?
d. Suppose conducting the study would cost $5,000. Will the government conduct the study? Explain your answer. (Assume they make decisions by maximizing expected monetary value.)

4. Casey has a phone that's worth $300. The probability that she'll damage it and need to replace it is 1/5. Casey is offered an insurance policy that costs $60 to replace her phone.

a. What is the expected monetary value of buying the insurance?
b. What is the expected monetary value of declining the insurance?

5. Zack has a laptop worth $1,000. Given his past, there's a 1/5 chance he's going to spill coffee on it and will need to replace it. At what cost for an insurance policy would Zack be indifferent between buying and declining the insurance?

6. In the example with the politician you may have wondered whether MEU would give a different recommendation if the politician had come to a different estimate for the probabilities of the states, even if the utilities stayed the same.

    - Are there any probabilities of the states where MEU would make the recommendation to End Search? If not, why not? If so, what are they? (Remember, the probabilites of the two states have to sum to 1.)
    - Are there any probabilities of the states where MEU would recommend to Extend Indefinitely? If not, why not? If so, what are they? (Remember, the probabilites of the two states have to sum to 1.)

7. Feeling free to build on your own example from a previous exercise, walk through an example where you use the multi-attribute approach to generate utilities for a decision matrix. Then make an assessment of the probabilities of the states, much like we did in the politician example. Finally, apply MEU to determine what it recommends. Your example should follow much like the politician one from this chapter. You should have at least three tables, two of which are part of the multi-attribute approach, and then a final table where you do the expected utility calculations. Make sure to explain your reasoning for how you constructed the tables. Remember that the states should be exhaustive and exclusive of the way the world could be (the things outside of your control), and the options should also be exhaustive and exclusive (the things in your control that you want to choose between).

6

Arguments about MEU

We've been developing a theory of decision making called Maximize Expected Utility (MEU). As with the development of any theory, it is a good idea to take a step back and consider: i) what is the domain/range/scope of the theory? ii) what are some reasons for thinking the theory is true? iii) what are some reasons for thinking the theory is false? This chapter will examine some preliminary answers to these questions.

6.1 The Domain of MEU

When we ask about the domain, range, or scope of a theory, we're asking what a theory is about. Number theory is a theory about numbers, it is not a theory about fashion, social justice, or economic systems. Jeremy Bentham's Utilitarianism, or Immanuel Kant's Deontology, are theories about ethical actions, they are not theories about the movement of celestial bodies, logical entailment, or the grammar of German.

Maximize expected utility is a theory of decision making. We need to take some care in what we take to count as a decision, such that MEU is a theory of it. For example, perhaps there are "choices" that we make that we don't think fall under the purview of decision making, and thus MEU isn't an analysis of that. Some examples that are, at best, fringe cases of decisions may include: when to fall asleep, being surprised at the sight of a snake, whether and when to hiccup, when to get hungry or dehydrated, etc.[1]

Relatedly, we need to also recognize that MEU is a mathematically precise theory of phenomena that are not typically so precise. That by itself is neither a point for nor against it. We have all sorts of examples where theories are more process than the real world applications require. Think of theories behind baking: it won't matter much whether your oven is 350 or 351 degrees for the cake to bake. What matters is that the theory (or in this case recipe) is an idealized

[1] Even these examples are a bit tricky, since we do seem to have some indirect way of nudging the times at which these things happen. E.g., if I eat at noon but then nothing else, there's a good chance I'm going to get hungry between 4pm and 7pm. But still, I don't really seem to choose the moment I get hungry.

version of something that is more approximate, and so that to the extent that the approximation is close enough, the theory gets the right answer or prediction.[2]

For the domain in question, there are all sorts of ways that ordinary people make decisions that aren't exactly like what MEU describes or prescribes, but are close enough. We almost never, for example, have in mind utility values that can be represented numerically. We also have mistaken beliefs about how the world works and our probability tables will not perfectly match the truth. Such imperfections are to be expected.

So what is MEU a theory of? It is a theory of choices that are made when we determine the approximate value of each relevant outcome, determine the approximate probability of each of these outcomes, and then use that information to estimate the expected value of each option. To the extent that choices are made in this way, is the extent to which MEU applies. Call these kinds of choices instances of **standard decision making**.

Standard decision making is closely connected with, and sometimes treated the same as, the concept of rationality. Rationality is a normative concept. As such, there is a three way distinction. There is the domain of the non-rational (or arational), the irrational, and the rational. The non-rational are those behaviors or decisions that we take to be outside the scope of rationality. This is the class of things where we think normative considerations are not applicable; it includes those things that we think are neither rational nor irrational. As an extreme example, the fact that unsupported pens fall to the ground is neither rational nor irrational - it is non-rational.[3] Similarly, the examples we listed above, like the very moment we fall asleep, whether or when we hiccup, etc., are not the sorts of behaviors we would ascribe rationality to, and hence are non-rational.

While many behaviors are non-rational, there is a set of behaviors for which we are willing to ascribe reasons, i.e., that can be scrutinized with an eye towards justifying (or failing to justify). That is the domain to which standard decision making applies and that MEU is a theory of. Here, when a choice is made with a good reason, that should be consistent with what MEU recommends. On the other hand, if a choice is made with a bad reason, then that should be at odds with what MEU recommends. To this end, we can understand MEU as a kind of theory of rationality (or, standard decision making).

There are interesting cases where it seems that choices are not made in the standard decision making way, and thus MEU does not apply. The philosopher L.A. Paul has argued that the decision to have a first child is one such example.[4] Many websites, life-coaches,

[2] We make all sorts of idealizations. Think of distances between cities: Is it really exactly 32 miles?

[3] Some authors use 'arational' to mean the same thing.

[4] See her paper "What You Can't Expect When You're Expecting" and her later book *Transformative Experiences*.

and family advisors give the impression that the decision to have a first child is a quintessential example of standard decision making that MEU is about: you are encouraged to think about the utilities of outcomes (how much satisfaction do you expect to get from being a parent?) and weight them by their probabilities (what are the chances that you will have a healthy child?) to estimate the expectations of the options to have a child or not. But L.A. Paul argues that contrary to such impressions, the decision to have a child is not like this at all. In fact, she argues that the decision to have a child is not a rational one. To be clear, she is not saying it is *irrational* to have a child (or not to). She is saying that it is a choice that is very different than the types of decisions that we make that MEU is intended to be a theory of. Or in our terminology: the choice to have a first child is a non-standard decision, and arguably non-rational.

Her argument boils down to the following. There are two types of *transformations* that can occur from the time before you have your first child to after. One is called an *epistemic* transformation. The claim here is that no one knows what it is like to experience having a child until they have had one. It doesn't matter if you've babysit a lot and taken care of children generally. The claim is that you simply do not have access to the relevant experience of having *your own* child. Perhaps you will like it, perhaps you will feel depressed, you simply don't know. Moreover, you have no guides to predicting what your experience will be like, no matter what experiences you have had so far - that's the idea behind the event being *transformative*. If you believe that epistemic transformations are possible and that having one's first child is an instance of it, then it is not possible for you to even get approximate utility values for the outcomes of your decision. So the choice of whether to be a child is not an instance of standard decision making.

The second kind of transformation is *personal*. In a personal transformation, an event changes you in a profound way such that your core preferences are affected. Having your first child is often said to produce personal transformations in some people, and often in unpredictable ways. If you believe that personal transformations are possible and that having your first child is an example of that, then again you have no way to use your current preferences as a guide for thinking about what your preferences will be in the outcomes of where you have your first child.

There is, L.A. Paul admits, a workaround that can convert the choice to have a child so that it comes really close to being a standard decision. The only thing we need to do in this workaround is not use your current preferences, but instead use only "external" considerations. Strictly speaking, the absence of using your current preferences

as guides for approximating utility values makes this workaround non-standard. Even if it is non-standard, we might still want to qual-ify it as a kind of "rational" deliberation - but let's leave that to the side for a moment. What matters here is that this non-standard ap-proach makes use of all the tools we have been describing as part of standard decision making. So what is this workaround?

Let's imagine there are four relevant groups that we will use as our external considerations: i) Lucky Parents, ii) Unlucky Parents, iii) Lucky Child-frees, and iv) Unlucky Child-frees. The Lucky Parents class includes those individuals that chose to have a child and it turns out that this outcome has an overall value that is higher than had they remained childless. The Unlucky Parents are those people who had a child but the value of this outcome is lower than had they remained childless. Similarly, the Lucky Child-frees is the class of individuals who remained childless and the value of that outcome is higher than if they had had a child. And finally, the Unlucky Child-frees are those individuals who remained childless but the value of this outcome is lower than had they had a child.

Using these four groups, we can ask what the empirical evidence is about their sizes. You could then use that information as a guide to estimate which group you're most likely to fall into. We'll assume, for the sake of illustration, that your preference is to be in the largest group.[5]

What does the empirical literature say? It is, as with a lot of sci-ence, a bit mixed and in some cases controversial. But we can say at least this much. There is little to no evidence that suggests that Lucky Parents is much larger than Unlucky Parents, nor is there evidence that Unlucky Child-frees is much larger than Lucky Child-frees. Moreover, there is evidence that suggests that overall well-being is likely to go down if you choose to have a child.[6]

The takeaway of L.A. Paul's argument is not about the rationality or irrationality of having a first child. Her argument is an illustration of what is and is not in the scope of a theory. She is arguing against a common conception that the choice to have a first child is an instance of standard decision making. As such an instance, there are either good justifications or poor justifications and the choice is either ratio-nal or irrational. But according to her argument, this way of thinking of the choice is mistaken. Instead, she's arguing that this choice is either non-rational or at best non-standard. In either case, her point is that the choice to have a first child is outside the scope of MEU.

For the rest of the chapter, we're going to narrow in on the domain of standard decision making and consider reasons for thinking MEU is a good theory of it.

[5] As we'll see later, this is probably not the right way to make the inference. What matters is the *relative* sizes of the groups. Reasoning about (causal) effectivness is tricky and we'll learn how to do it better in later chapters.

[6] See McClanahan and Adams (1989), Simon (2008), and Evenson and Simon 2005. Some reports say that fathers enjoy higher levels of life satisfaction, but mothers do not (Nelson et al. 2013).

6.2 Long Run Arguments for MEU

Let's assume that we're considering any example of standard decision making. Some of the most common arguments that MEU is a good theory of such decision making are called "long run" argument. Long run arguments claim that you will be better off in the long run if you make decisions by maximizing expected utility. This argument is based on the law of large numbers. Suppose that an outcome has a probability p of occurring, e.g., the chance of rolling a one with a fair six-sided die is $1/6 = 0.167$. Then the law of large numbers says that as the number of rolls (or 'random experiments') n increases towards infinity, the proportion of rolls that lead to one will be 0.167. If you roll only a few times, you might never see one, or you might see it a high number of times, but as you roll the die hundreds of times, the fraction of times you see one relative to the other numbers becomes increasingly closer to $1/6$.

When we apply this idea to decision making, we are saying that if you face the same decision over and over again and use the maximize expected utility strategy, then the utilities you will get on average will converge to the maximum expected utility. For example, suppose that you have a choice between getting \$1 for sure, or getting \$3 if a fair coin lands heads and nothing if it lands tails. The expected utility of the coin toss choice is \$1.5, which is higher than the expected \$1 of the 'for sure' choice. If you only get to flip the coin once, then there's a 50% chance you'll get nothing at all. But the law of large numbers says that if you get to flip the coin many times, on average you'll get \$1.5. So while choosing the 'for sure' option 100 times will lead to \$100, choosing the 'flip coin' option 100 times will, on average, yield \$150. For this reason, MEU recommends the 'flip coin' option.

How good is the long run argument? There is a famous phrase from Keynes: "in the long run we're all dead."[7] Let's unpack how this could be turned into an objection against the long run argument. One way of thinking about the law of large numbers is that we have to do infinitely many experiments (e.g., coin flips) for the average payoff to converge to the expected utility. Since we don't actually have an infinite amount of time, one could say that the use of the mathematical result shouldn't be used in this context.

There are, however, many places where mathematical idealizations don't map onto the world but are nevertheless good enough approximations. We need not get bogged down here. The long run argument needn't necessarily require that infinitely many experiments are done. It is sufficient to notice that over time the amount of variation is expected to decrease and there is a tendency to converge towards the expected utility. If you flip a coin twice, there are four

[7] Around the same time as Keynes, Frank Ramsey had the following to say: "In time the world will cool and everything will die; but that is a long time off still, and its present value at compound discount is almost nothing. Nor is the present less valuable because the future will be blank. Humanity, which fills the foreground of my picture, I find interesting and on the whole admirable. I find, just now at least, the world a pleasant and exciting place. You may find it depressing; I am sorry for you, and you despise me. . . .On the other hand, I pity you with reason, because it is pleasanter to be thrilled than to be depressed, and not merely pleasanter but better for all one's activities."

possible sequences: HH, HT, TH, and TT. In two of the four we get 50% heads, in one we get 100% heads, and in the last one we get 0% heads. As we increase the number of flips, there is only one way to get 100% heads, but the number of ways to get a sequence of around 50% heads increases.[8]

Perhaps a different way of making the objection is that for many of our decisions we will only make them once or a few times. Whether or not to go to war or whom to marry are decisions that happen infrequently enough that the law of large numbers doesn't seem to apply (leaving aside considerations about whether these are instances of transformative experiences, as discussed above). In fact, some of our decision making behaviors seem to indicate that when we are faced with uncertainty in a decision we will only make once, we prefer *not* to go the route of maximize expected utility but rather with an option where we are more certain about the outcome. We'll see this point when we talk about risk aversion.

6.3   Two Kinds of Arguments Against MEU

There are two styles of argument against the strategy of maximizing expected utility. One style of argument against MEU is to show that its assumptions can be used to generate paradoxes, i.e., conclusions that are at best unintuitive and at worst contradictory. Conclusions that are unintuitive are not necessarily surefire objections - one can always "bite the bullet" and accept that some consequences are quite surprising. But at the very least one must accept that these are "costs" of accepting the position. To be clear, these arguments we're going to consider target the Maximize Expected Utility principle as a normative theory. That is, we'll look at examples where we're considering how we think that an ideal agent should reason, and show that MEU is not consistent with such an ideal agent.

The Maximize Expected Utility principle is sometimes interpreted as a normative rule. As such, failing to behave in accordance with it is evidence of non-rationality or even irrationality (assuming the behavior is within the scope of MEU). MEU can also be interpreted as a description of how reasoning is done, at least when it's done at its best. In this case, if we make sure to set up situations where decision making hangs only on the variables in question, but agents fail to make decisions as predicted by MEU, then that is evidence that MEU is false.

Evaluating MEU from a descriptive standpoint can be a bit tricky and it's easy to slide back and forth between the normative and descriptive divide. Here is a helpful example.[9] Suppose some engineers come up with a principle about building bridges. If a bridge col-

lapses, it is possible to lay blame on the construction of the bridge, not on the principle, e.g. perhaps the joints weren't secured properly and that's what caused the collapse. On the other hand, if a bridge is built with all the appropriate specifications and counts as a good representation of a bridge, and yet it still collapses, that can be treated as evidence that there's something wrong about the principle governing bridge building. In the same way, by carefully constructing experiments so that humans are good representations of "rational decision making", we can test whether MEU is a good description of it.

So we are also going to look at a few popular examples of how human reasoning appears to be at odds with MEU from a descriptive perspective. In order to keep things simple, we'll work with monetary amounts. As we've seen, the utility of money has a diminishing effect. As long as we keep this caveat in mind, it's simpler to represent the examples with money. But rest assured that the actual experiments will have controlled appropriately for some of these complications.

6.4 Arguments Against Normative MEU

6.4.1 Allais Paradox

The Allais Paradox is named after its discoverer, Maurice Allais, a Nobel Prize winning economist. Consider the four following gamble options in a lottery that has exactly 100 tickets.

	Ticket 1	Tickets 2-11	Tickets 12-100
Gamble 1	$1M	$1M	$1M
Gamble 2	$0	$5M	$1M
Gamble 3	$1M	$1M	$0
Gamble 4	$0	$5M	$0

Some decision theorists have suggested that it is perfectly reasonable to prefer Gamble 1 over Gamble 2. Even though Gamble 2 has an expected utility of $1.39M, if ticket 1 is drawn you get nothing at all. Gamble 1, on the other hand, makes winning $1M a sure thing. If you only get to play this lottery once and you're averse to taking risk, Gamble 1 seems the way to go.

In addition, those same decision theorists say it is perfectly reasonable to prefer Gamble 4 over Gamble 3. Gamble 3 has an expected utility of $0.11M while Gamble 4 has an expected utility of $0.5M.

As a matter of empirical fact, people tend to prefer 1 over 2, and 4 over 3. But even if people didn't, some argue that an ideal agent with these preferences is still rational.

Here's the problem: There is no utility function that can produce the result that Gamble 1 ≻ Gamble 2 and Gamble 4 ≻ Gamble 3. We can actually prove this. To do so, we need to show that the difference in expected utility between Gamble 1 and Gamble 2 is the same as the difference between Gamble 3 and 4. If that difference is the same, then there's no way that Gamble 1 ≻ Gamble 2 and Gamble 4 ≻ Gamble 3 (although there's the limiting case where they are equal, but that would mean indifference, which is not what we're after). Note that the probability of drawing ticket 1 is 0.01, the probability of drawing one of tickets 2-11 is 0.1, and the probability of drawing one of tickets 12-100 is 0.89. Also, keeping in mind that utility doesn't correspond directly to money, let's say u is a utility function that takes money as input (like in our discussion of marginal utilities).[10] If we plug these probabilities into the expected utility equations, we can calculate the following differences:

$$Exp(\text{Gamble 1}) - Exp(\text{Gamble 2})$$
$$= u(1\text{M}) - [0.01u(0\text{M}) + 0.1u(5\text{M}) + 0.89u(1\text{M})]$$
$$= 0.11u(1\text{M}) - [0.01u(0) + 0.1u(5\text{M})]$$

$$Exp(\text{Gamble 3}) - Exp(\text{Gamble 4})$$
$$= [0.11u(1\text{M}) + 0.89u(0)] - [0.9u(0\text{M}) + 0.1u(5\text{M})]$$
$$= 0.11u(1\text{M}) - [0.01u(0) + 0.1u(5\text{M})]$$

Note that final lines of the equations show that the differences between Gamble 1 and Gamble 2, and Gamble 3 and 4, are exactly the same. That means that no matter what utility function you give for money, there's simply no way to make Gamble 1 have a higher expected utility than Gamble 2 and simultaneously have a higher expected utility for Gamble 4 than Gamble 3.

It's worth pointing out that the situation we are working in is a decision under risk. That is, this is a situation where, although we are uncertain about which outcomes will occur, we can at least assign probabilities to those outcomes. What the Allais Paradox highlights, at the very least, is that we tend to be (and, we think ideal agents should be) averse to risk even when we know what those risks are.

The next paradox will show that our aversion is not only to known risks, but also to unknown risks.

6.4.2   Ellsberg's Paradox

The Ellsberg's Paradox is named after Daniel Ellsberg, who discovered it while being a Ph.D. student in economics at Harvard in the 1950s.

[10] One might think that there are some assumptions that we have to make for this to work, e.g., u has to be a monotonic function, i.e., we cannot allow that utilities go down while money goes up. But as we'll see momentarily, we don't even need this kind of assumption.

In this example, let's suppose that there is an urn filled with 90 balls, numbered 1 through 90. We'll assume that balls 1-30 are red. Further, we'll assume that balls 31-90 are either black or yellow, but we leave it unknown what the proportion between black and yellow is. One ball will be blindly drawn from the urn. Given this setup, consider the following gambles and their payouts given the color of the ball that will be drawn.

	Balls 1-30	Balls 31...	...through 90
	Red	Black	Yellow
Gamble 1	100	0	0
Gamble 2	0	100	0
Gamble 3	100	0	100
Gamble 4	0	100	100

Along the same lines as the Allais paradox, if we're averse to risk, it seems reasonable to prefer Gamble 1 over Gamble 2. That's because, even though drawing a red ball is not guaranteed (there's a 30/90 = 33% chance of drawing red) we at least know what that risk is, whereas the probability of drawing a black ball is unknown (except that it's somewhere between 0/90 and 60/90). It will turn out, however, even if you believe that there are more black balls than red balls (which means your odds of picking a black ball is higher than a red ball), we'll still get the paradox. For simplicity's sake, let's assume Gamble 1≻Gamble 2.

If we look at Gamble 4 we notice that the chances of winning $100 is 60/90 because, even though it is unknown what proportion of black to yellow balls there are, we get $100 whether the ball is black or yellow (but nothing if it's red). Gamble 3, on the other hand, has a 30/90 chance of paying out $100 (when red is draw) *plus* an unknown chance of paying out $100 if the ball is yellow. For all you know, this chance could be 0/90 and upwards of 60/90. Indeed, if all the balls from 31 through 90 are yellow, then Gamble 3 would pay out $100 with a 100% chance. But again, you don't know the proportion of black to yellow. So, if we consider risk aversion again, it seems that Gamble 4≻Gamble 3.

The problem is that there is no utility function that can recommend both Gamble 1≻Gamble 2 and Gamble 4≻Gamble 3. We can demonstrate this by showing that Gamble 2≻Gamble 1 if and only if Gamble 4≻Gamble 3. To demonstrate this, we calculate the difference in expected utilities between Gambles 1 and 2, and Gambles 3 and 4. For simplicity, we'll use M to represent the utility of $100 and assume that the utility of $0 is 0. We'll use B to represent the number of black balls.

$$Exp(\text{Gamble 1}) - Exp(\text{Gamble 2})$$
$$= 30/90\text{M} - \text{B}/90\text{M}$$
$$= 30\text{M} - \text{BM}$$

$$Exp(\text{Gamble 3}) - Exp(\text{Gamble 4})$$
$$= 30/90\text{M} + (60 - \text{B})/90\text{M} - 60/90\text{M}$$
$$= 30\text{M} + (60 - \text{B})\text{M} - 60\text{M}$$
$$= 30\text{M} - \text{BM}$$

Notice that the differences in expected utilities in the pairs of gambles are identical. That means there's no way for Gamble 1≻Gamble 2 and simultaneously Gamble 4≻Gamble 3. Moreover, if we were to insist that Gamble 2≻Gamble 1, then that will force us to commit to Gamble 3≻Gamble 4; a situation where we are avoiding known risks and preferring unknown risks.

The Ellsberg and Allais paradoxes have the same flavor, but they illustrate a subtle difference. Recall that in addition to decisions under certainty, we distinguished between two different kinds of *un*certainty: known risk (i.e., decisions under risk) and unknown risk (i.e., decision under ignorance). The Allais paradox is an illustration of our preference for options that are certain over options with known risk. The Ellsberg paradox illustrates that we prefer known risk to unknown risk. The Maximize Expected Utility principle, however, does not distinguish between these three kinds of options. It treats certainty, known risk, and unknown risk as all the same - all that matters is the expected utility of the option.

What the Ellsberg and Allais paradoxes have in common is that they invite us to violate a rule known as the **Sure Thing Principle**, which we discuss in the next section.

6.4.3   *The Sure Thing Principle*

Consider again the table from the Allais paradox. Notice that Gambles 1 and 2 have the same outcome for tickets 12-100. Notice also that Gambles 3 and 4 agree on tickets 12-100.

	Ticket 1	Tickets 2-11	Tickets 12-100
Gamble 1	**$1M**	**$1M**	*$1M*
Gamble 2	**$0**	**$5M**	*$1M*
Gamble 3	**$1M**	**$1M**	*$0*

	Ticket 1	Tickets 2-11	Tickets 12-100
Gamble 4	**$0**	**$5M**	*$0*

The **Sure Thing Principle** says that when you compare two options, like Gambles 1 and 2, you can ignore the states on which they agree (in italics) and compare them only on the states in which they disagree (boldfaced). In this case, the Sure Thing Principle says that when comparing Gambles 1 and 2, we don't need to consider tickets 12-100, we can just look at the states of ticket 1 and tickets 2-11. The same goes for comparing gambles 3 and 4.

If we ignore the last column, as the Sure Thing Principle says we can, we'll notice that the comparison between gambles 1 and 2 is *exactly* the same as the comparison between gambles 3 and 4. That is, from the perspective of ticket 1 and tickets 2-11, Gamble 1 is the same as Gamble 3, and Gamble 2 is the same as Gamble 4.

The Maximize Expected Utility rule entails the Sure Thing Principle. That is, any violation of the Sure Thing Principle will also be a violation of Maximize Expected Utility. Or put differently again, if you like MEU, then you also should like the Sure Thing Principle. In the above Allais table, the Sure Thing Principle effectively says that the comparison between gambles 1 and 2 is the same comparison as gambles 3 and 4. By these lights, the recommendation by MEU is correct. That means that our intuition that Gamble 4≻Gamble 3 but Gamble 1≻Gamble 2 is mistaken. The Allais paradox, according to this analysis, is starting off on the wrong foot.

A similar analysis of the Ellsberg paradox tells us that our intuition that Gamble 4≻Gamble 3 and Gamble 1≻Gamble 2 must be mistaken. In this case, the state that the Sure Thing Principle says we can ignore is when the ball is yellow. By doing so, we see that the comparison between gambles 1 and 2 is the same as the comparison between gambles 3 and 4.

Given that our initial intuitions about the gambles are at odds with the Sure Thing Principle, it is perfectly reasonable to ask which one we should adopt. Should we trust our intuitions, or should we trust the Sure Thing Principle? When addressing the Allais Paradox, Leonard Savage had this to say in favor of the Sure Thing Principle:

> if one of the tickets number from 12 through 100 is drawn, it does not matter, in either situation which gamble I choose. I therefore focus on the possibility that one of the tickets numbered from 1 through 11 will be drawn, in which case [the choice between Gamble 1 and Gamble 2 and between Gamble 3 and Gamble 4] are exactly parallel . . . It seem to be that in reversing my preference between [Gamble 3 and Gamble 4] I have corrected an error. (Savage 1954: 103)

Not everyone is convinced, however. In fact, in Savage's axiomatization of the expected utility principle, the most controversial axiom is the Sure Thing Principle. At the very least, one thing we can say is that neither the Allais and Ellsberg paradoxes are knock-down arguments against Maximize Expected Utility. Another thing we can say is that we've uncovered an important assumption about MEU, namely that if it's true, the Sure Thing Principle has to be true as well.

6.4.4 St Petersburg Paradox

Consider the St. Petersburg game. The game starts by flipping a fair coin. If it lands tails you flip the coin again. When it lands heads up the game is over. The idea is to play for as long as you can. There's a prize worth 2^n units of utility, where n is the number of times the coin was flipped. So, if your first flip was heads, then the prize would be 2 units of utility. If you flip the coin 4 times where you get three tails and then a heads, i.e., TTTH, then the prize is $2^4 = 2 \times 2 \times 2 \times 2 = 16$. How much would you be willing to pay to play the St. Petersburg game?

According to MEU, you should be willing to pay any finite amount of utility to play the St. Petersburg game. Here's why. The probability of flipping a coin n times, where n is the first time that the coin lands heads, is $(\frac{1}{2})^n$. The payoff for flipping the coin n times is 2^n. The expected utility rule says to weight the utility of each possible outcome according to its probability, and then sum them up. This gives us the following sequence of possible series of flips:

$$\text{H} + \text{TH} + \text{TTH} + \cdots = \left[\frac{1}{2} \times 2\right] + \left[\frac{1}{4} \times 4\right] + \left[\frac{1}{8} \times 8\right] + \cdots$$
$$= 1 + 1 + 1 + \cdots$$
$$= \infty$$

In other words, the expected utility of the St. Peterburg game is infinitely many utilities. So, we should be willing to pay any amount that is less than infinitely many utilities. Most people think it is absurd to pay any large finite amount. It's unlikely that you're willing to pay even a few hundred units of utility. The probability of getting a large sequence of tails before a heads is diminishingly small.

This paradox was discovered by Daniel Bernoulli (1700-1782), a Swiss mathematician that was working in St. Petersburg at the time. In response to the paradox, some, like Buffon did in 1745, argue that some outcomes should be ignored if they are beyond reasonable or practical concern, if they are "morally impossible".

There is a closely related idea to Buffon's suggestion, known as *de*

minimis risk. In risk analysis, sufficiently improbable outcomes, like a comet strike, are ignored because their probabilities are too low to play a significant role in risk analysis.

If we ignore sufficiently improbable events, then we can block the St. Peterburg paradox. Say that we only consider outcomes that have at least a 5% chance of happening (5% is actually still quite high from a risk analysis perspective, we're just using this number to illustrate a point). Then the last outcome you would consider in the sequence of possible series of coin flips is TTTH, which has a probability of $0.5^4 = 0.0625$ or a 6.25% chance. In other words, you would ignore all sequences where there are 5 or more flips. That is, you would only consider the follow sequence:

$$\frac{1}{2} \times 2 + \frac{1}{4} \times 4 + \frac{1}{8} \times 8 + \frac{1}{16} \times 16$$
$$= 1 + 1 + 1 + 1$$
$$= 4$$

There are two related concerns with this solution. One, we haven't provided any reason for why 5% should be the minimum probability that an outcome needs to have to be considered. Any such line is likely to be ad hoc. Two, and more generally, why would it be rational to ignore highly improbably outcomes, especially if the corresponding utilities have large magnitudes?

Another resolution to the paradox is to try to put an upper limit on the utility scale of the decision maker. It seems reasonable to accept that utilities are bounded, that utilities for an individual cannot grow indefinitely. That is, there might be some kind of "hedonic plateau".[11] To see how this idea can be used to resolve the St. Peterburg paradox, suppose that L is the finite upper limit of utility. Then at some point in the sequence of possible outcomes the amount of utilities gained no longer grow, even though the probabilities continue to get smaller:

$$\left[\frac{1}{2} \times 2\right] + \left[\frac{1}{4} \times 4\right] + \left[\frac{1}{8} \times 8\right] + \cdots + \left[\frac{1}{2^k} \times L\right] + \left[\frac{1}{2^{k+1}} \times L\right] + \cdots$$

We can think of the sequence of consisting of two parts. The first part of the sequence consists of all the utilities before the upper bound L is reached. Let's say k is the position where the upper bound L is first reached. Then the first part of the sequence ends at $k - 1$. Since we are taking the sum of products, we can express it succinctly as:[12]

$$\sum_{i=1}^{k-1} (1/2)^i \times 2^i$$

Note that this sum is finite.

[12] I know, we said we'd keep things not-so technical. It just takes up so much page space! If you're not familiar with this notation, see this explanation.

The second part of the sequence consists of all the utilities where the upper bound L is reached. We can express this as:

$$\sum_{i=k}^{\infty}(1/2)^i \times L$$

Notice that even though this sequence continues indefinitely and is infinite, each addition later in the sequence becomes smaller than the previous one, approaching zero in the limit. Hence, this sum will also be finite.

The result of adding the finite sum from the first part of the sequence that's below the upper bound L and the finite sum of the second part of the sequence is itself going to be finite. Let's say that amount is C. The amount you should be willing to pay to play the St. Petersburg game is now anything up to the amount C, but no more. We thereby avoid the paradox.[13]

If the first resolution to the paradox is ad hoc, where we use the idea of *de minimis* risk, then this second resolution where we introduce an upper bound to utilities also seems ad hoc. Moreover, the paradox can be reformulated without requiring that the sequence of flips is infinite. It is enough to recreate the paradox just in case the expected utility of the game is unreasonably high in comparison to how much we think it is reasonable to pay to play.

There are other angles to approaching the St. Peterburg paradox. Richard C. Jeffrey, for example, pointed out that whoever is offering the game is committing themselves to possibly paying the player an indefinite amount of money. But no one has such an indefinitely large bank and thereby cannot possibly fulfill such a commitment. This means that the assumptions of the game cannot be valid.

Note that the success of Jeffrey's response depends on there being some limit on what is possible to pay out to a player. If a bank has access to a Nozick-style experience machine, they might offer payouts in the form of intensely happy experiences. Here too, however, we can ask whether there is some natural upper bound on utilities. If there is, then that would not only support Jeffrey's solution, but it would also be a way to address the 'ad hoc' objection given to the second resolution above (where we posited some upper bound L on utilities).

[13] This solution goes back to the 19th century mathematician Cramer and was also endorsed by the Nobel Prize winner Kenneth Arrow.

6.4.5   The Two Envelope Paradox

(Intentionally omitted - this is bonus material.)

6.5   Arguments Against Descriptive MEU

The Maximize Expected Utility principle is sometimes interpreted as a normative rule. As such, failing to behave in accordance with it is evidence of non-rationality or even irrationality (assuming the behavior is within the scope of MEU). MEU can also be interpreted as a description of how reasoning is done, at least when it's done at its best. In this case, if we make sure to set up situations where decision making hangs only on the variables in question, but agents fail to make decisions as predicted by MEU, then that is evidence that MEU is false.

Evaluating MEU from a descriptive standpoint can be a bit tricky and it's easy to slide back and forth between the normative and descriptive divide. Here is a helpful example. Suppose some engineers come up with a principle about building bridges. If a bridge collapses, it is possible to lay blame on the construction of the bridge, not on the principle, e.g. perhaps the joints weren't secured properly and that's what caused the collapse. On the other hand, if a bridge is built with all the appropriate specifications and counts as a good representation of a bridge, and yet it still collapses, that can be treated as evidence that there's something wrong about the principle governing bridge building. In the same way, by carefully constructing experiments so that humans are good representations of "rational decision making'', we can test whether MEU is a good description of it.

In this section we're going to look at just a few popular examples of how human reasoning appears to be at odds with MEU. In order to keep things simple, we'll work with monetary amounts. As we've seen, the utility of money has a diminishing effect. As long as we keep this caveat in mind, it's simpler to represent the examples with money. But rest assured that the actual experiments will have controlled appropriately for some of these complications.

6.5.1   Risk Aversion

We have seen the topic of risk aversion in the examples of the Allais and Ellsberg paradoxes. There, it was assumed that an ideal agent prefers options with less risk over options with more risk, even if the expected utilities of the options are identical. It turns out that humans appear to behave the same way.

Consider the following decision matrix that describes payouts for a fair coin flip.

	Heads	Tails	Expected Utility
Option A	$50	$50	$50

	Heads	Tails	Expected Utility
Option B	$0	$100	$50

When given these options, most people will have a preference for Option A. According to MEU, however, we should be indifferent between A and B. The reason we should be indifferent is that the expected utilities for both options is $50.

MEU does not take into account the range of possible outcomes, except insofar as they contribute to the calculation of the expected utility. For example, we can increase the expected utility of Option B by increasing the payout for Tails. But so long as we increase all the payouts for option A to match the increased expected utility of B accordingly, the two options will be equally preferred by MEU. Humans tend not to be indifferent, however.

In fact, up to a certain point, it seems humans prefer a lesser expectation if it is more certain than a higher expectation that is less certain. For example, in the table below Gamble 1 has an expectation of $2,400 for sure, while Gamble 2 has an expectation of $2,409. While MEU recommends Gamble 2, humans typically prefer Gamble 1.

	Tickets 1-33	Tickets 34-99	Ticket 100
Gamble 1	$2,400	$2,400	$2,400
Gamble 2	$2,500	$2,400	$0

To make the case that it really is risk aversion that seems to be driving the decision making here, as opposed to something about utility functions, experimenters can inject some risk into the decision but keep the utilities relatively stable. For example, Gamble 3 is just like Gamble 1 except now there is uncertainty in the form of known risk: tickets 34 through 99 do not have a payout. The expected utility of Gamble 3 is $816. We make the same kind of change to Gamble 2 in order to create Gamble 4, which now has an expected utility of $825. When Gamble 3 is compared to Gamble 4, now people typically prefer Gamble 4 over Gamble 3.

	Tickets 1-33	Tickets 34-99	Ticket 100
Gamble 1	$2,400	$2,400	$2,400
Gamble 2	$2,500	$2,400	$0
Gamble 3	$2,400	$0	$2,400
Gamble 4	$2,500	$0	$0

If the table containing Gambles 1-4 seems familiar, it's because it has the same structure as the Allais paradox. The main difference here is that in the case of the Allais paradox we were concerned with whether an idealized rational agent would have preferences: Gamble 1 ≻ Gamble 2, and Gamble 4 ≻ Gamble 3. Here we are evaluating MEU as a descriptive theory, so we are comparing what MEU predicts against what the empirical data says. MEU predicts that both Gamble 2 ≻ Gamble 1, and Gamble 4 ≻ Gamble 3. According to a classic experiment done by Kahneman and Tverksy, 82% of participants preferred Gamble 1 over Gamble 2, and using the same experimental participants, 83% preferred Gamble 4 over Gamble 3. These findings have been reproduced by many researchers all over the world. This seems to be a clear mark against MEU as a descriptive theory.

6.5.2   Loss Aversion

In addition to being averse to risk, humans tend also be averse to loss. To illustrate this, we can set up a situation where the only two options are both risky. However, in one of the options, there's a possibility of actually losing some utilities. In order to compensate for this possible loss, the payout for the "winning'' situation is increased by the same amount. Here's an example.

	Heads	Tails	Expected Utility
Option A	$0	$100	$50
Option B	-$10	$110	$50

Notice that the expected utilities of the two options are identical, so MEU does not have a recommendation (i.e., we should be indifferent). People, however, tend to prefer Option A over Option B because in the latter they have to actually pay $10 if the coin lands Heads, whereas they don't lose any money if they go with Option A. And yes, that preference holds even though Option B would have a $10 higher payout than Option A if the coin landed Tails.

6.5.3   Endowment Effect

Suppose half the students in the class are given a mug. Not all the students that got a mug really need it, and many of the students that didn't get a mug could really use one, and lots of students will be indifferent to mugs. In order to try to sort out the situation, we set up a little marketplace for people to exchange things. Before the mugs

go into the marketplace, however, we ask every student what they think the dollar value of a mug is.

On the "classical view" of MEU, it shouldn't matter whether someone was a mug recipient or not when they provide an assessment of a mug's value. Of course we expect there to be some variation among students in how much they think a mug is worth. But what we don't expect is that there is a systematic pattern in the dollar amounts they give that reliably partitions the class into those that received the mugs and those that didn't.

And yet, that's exactly what we tend to find. Students who are given a mug will tend to think that a mug is worth a higher amount than those who were not given a mug. This is known as the *endowment effect*. Humans will more highly estimate the value of something simply because they possess it than if they were not in possession of it. There's a good chance that if you've bought and sold a car you may have experienced this already: when you own the car, you are likely to think its value is higher than if you're the one trying to buy the car.

Here's another way of illustrating the endowment effect. Suppose Option A is that you get no money and then flip a coin, while Option B is that you are given $100 and then flip a coin. The payouts for the options given the coin flips are described in the following table.

	Heads	Tails	Expected Utility
A: Given $0 and flip	$0	$100	$50
B: Given $100 and flip	-$100	$0	$50

The expected utilities for both options are exactly the same ($50), and so according to MEU we should be indifferent between A and B. However, empirical data shows that humans have a preference for option A. Even though we stand to gain the same amount with either option, we seem to dislike having to give up something we are endowed with. In other words, we seem to prefer maintaining our utilities with the possibility of increasing them over the option to increase them with the possibility of losing them. In some ways, the endowment effect combines our aversion to risk and loss.

6.5.4   Prospect Theory

A common theme between risk aversion, loss aversion, and the endowment effect is that they are reference dependent: people's decision making seems to depend on a reference point rather than from some absolute perspective or a "view from nowhere". That is, when people make a decision, they take into account information about

their current situation. If we were to think of these decisions in absolute terms, as MEU assumes that we do, then all that is relevant is the position we can expect our future selves to be in after the decision. For example, in the endowment effect, both option A and B have the same future expectation: on average I can expect to be $50 richer. And yet, when I actually do my decision making, I seem to also consider my past and present situation, in addition to the future.

Prospect Theory was developed by Kahnemann and Tversky[14] to explain these and other related effects of human decision making. The main idea is that decision making happens in two stages. In the first stage, there is a process called *editing*. Here, the outcomes of a decision are ordered using some heuristics ("rules of thumb'"). For example, one heuristic to decide how dangerous or risky an action might be is to assess how easy it is to think of examples of where that action goes wrong. An example that is commonly brought up is comparing the dangers of air travel vs automobile travel. Airplane crashes are typically horrific and brought up in the news, making it easy for us to recall such examples. But car accidents are typically not headline worthy, even if they are more prevalent. That's the idea behind salience: by any statistical measure, the option to drive across the United States is more dangerous than the option to fly across it, and yet many people at least *feel* like flying is more dangerous because it is easier to recall the negative outcomes of flying. Kahnemann and Tversky, along with many researchers since, have explored many of the systematic ways humans tend to rely on heuristics, as well as biases, to help simplify decision making.

The second stage of human decision making, according to Prospect Theory, is called the *evaluation* phase. Here people combine information about the probabilities of outcomes with the value or utility of those outcomes - much like we've described in calculating expected utilities. However, unlike standard expected utility theory, the expected utility function in Prospect Theory passes through a reference point. In a similar way that money has a diminishing marginal utility, so does the Prospect utility function, except when losses are being considered, the S-shaped function is asymmetrical. That is, losses will have lower absolute values of utility than their gains counterpart. For example, gaining $0.05 might have utility 17, but losing $0.05 would have a utility of -40.

To what extent is Prospect Theory an alternative to Maximize Expected Utility? It depends in part how we characterize the relationship between the theories. Here is an optimistic characterization.

Notice that in the second stage, the evaluation phase, Prospect Theory claims that decisions makers run the outcomes through a process where they weight the probability of the outcome with the

[14] (CITE 1979 Prospect Theory)

utility assigned to the outcome. That is exactly what MEU also does. If we are minimalistic about the commitments of MEU, we could say that MEU is not committed to how utilities are assigned and whether they go through a reference point or not. That is, one could reasonably argue that the main thrust of MEU is that we compare options by taking the sum of the weighted probabilities (or equivalently, the weighted utilities) of their outcomes.

In other words, if maximize expected utility makes no commitments about the nature of the utility function, but Prospect Theory does, then Prospect Theory is a refinement of MEU.

Not everyone will agree with this characterization of the relationship between MEU and Prospect Theory. But at the very least we can say that there is a way for MEU to salvage itself as a descriptive theory, at least for the observations we've concern ourselves with so far. The next result, however, is difficult for MEU to salvage itself from.

6.5.5   Weighting Effects and Why They Matter

Consider the following prospects.

- Prospect 1: $6,000 with 45%

- Prospect 2: $3,000 with 90%

- Prospect 3: $6,000 with 0.1%

- Prospect 4: $3,000 with 0.2%

In a classic experiment of 66 participants, 86% said they preferred Prospect 2 over Prospect 1, whereas 73% preferred Prospect 3 over Prospect 4. From the perspective of MEU, this is puzzling because Prospects 1 and 2 have the same expected utility, as do Prospects 3 and 4. Moreover, the relationship between 1 and 2 is that Prospect 2 has twice the chance of getting half the amount in Prospect 1, and that is also true of how 4 relates to 3. So at the very least we would expect that if people prefer Prospect 2 over 1, then for the same reason theoretical considerations would lead us to predict a preference for 4 over 3. And yet, that's not what we find empirically.

What explains why people prefer 2 over 1, but 3 over 4? The most agreed upon hypothesis is that we are typically poor at reasoning with low probabilities. While it is relatively straightforward to perceive that 90% is twice as much as 45%, the percentages of Prospects 3 and 4 are so low that we don't pay much attention to their relative sizes. Sure, we might think, strictly speaking 0.2% is twice as high as 0.1%, but both are so low in terms of chances of winning that we treat them as being similar enough. Contrast that with the comparison of $6,000 and $3,000, where it is very easy to appreciate that one

is twice the amount as the other. In brief, our inability to distinguish or appreciate differences in small probabilities is thought to explain what goes wrong here.

Prospect Theory takes seriously this feature of human reasoning. On the one hand, people typically underestimate moderate and large probabilities, but overestimate small probabilities. To incorporate this in a theory of decision making, Prospect Theory adds a weighting function to the probability function. That is, whereas MEU takes probabilities at their face value, the weighting function will manipulate lower probabilities to make them seem higher than they are, and make moderate and high probabilities seem lower than they are.

In our characterization of the relationship between MEU and Prospect Theory, we said that MEU might not be committed to a particular utility function. That is, there doesn't seem to be anything inconsistent with letting MEU be open to the kind of asymmetric function that Prospect Theory proposes. The same cannot be said when it comes to probabilities. Here Prospect Theory seems to be making adjustments that are at the very least in tension with the spirit of MEU. The tension can be seen by asking ourselves, are humans making a mistake when they are reasoning with probabilities?

We have to be very careful here about how we think of the word 'mistake'. It is not intended to signal a normative-descriptive divide. Our bridge building example will be helpful here. When a bridge collapses, we can try to figure out what the cause was. Was there a mistake in the design of the bridge, in which case the cause of the collapse was in the engineering (the theory)? Or was there a mistake in the construction of the bridge, in which case the cause of the collapse was in the application of the design or poor materials (the experiments or observations).

Prospect theory says that the cause of the collapse was in the engineering. What humans do is systematic and should be taken as an expression of rationality. We need to therefore update our theory of rational decision making in order to make it account for humans. MEU, on the other hand, says that the cause of the collapse was in the application. When humans are tired or drunk, for example, they don't always choose the same options as they would when they are clear and sober. There are lots of instances where people are arational or irrational. MEU sees weighting effects as an example of *not* expressing rationality.

In short, Prospect Theory says to take the experiments seriously and update the theory. MEU says to take the theory seriously and recognize that the experiments are done when humans are not at their best, or even approximations of it. What should we decide?

Let us take a step back and consider what we want out of a the-

ory of decision making. There are at least two things we might care about. On the one hand, we might care about making predictions about how humans make decisions. On the other hand, we might care about having a theory that we can use as a guide to making decisions.

With respect to making predictions, it seems that Prospect Theory is superior to its ancestor, MEU. In fact, Prospect Theory has done so well that it has spawned an entire cottage industry of researchers that study heuristics (or "rules of thumb") in decision making. In turn, these heuristics can be used to nudge people in their decisions. Sometimes this is for the better, sometimes for the worse. Here is an example of the better, assuming that being an organ donor is a good thing from society's perspective. The driver's license office can offer you the same choice (to be an organ donor or not) in two different ways: you opt in to be a donor, or you opt out of being a donor. Notice that you are just as free to make the choice either way, but there is a certain amount of "work" you have to do for one of the options, and which option that is differs in the two contexts. In the context where the organ donor option is the default and people can opt out, more people are likely to be an organ donor than if they had to do the work of "opting in" to be one.

As you can imagine, however, once it can be predicted how people make decisions, this can also be used in ways that will not necessarily benefit us. Advertising is the primary example of how our decisions are influenced by interests that may not be in line with our own.

Maybe the engineers are right though, and what we need to do is become better bridge builders. To that end, Prospect Theory is not necessarily a theory that we want, if what we want is a theory that helps guide our decision making in a way that improves our chances of fulfilling our interests or preferences. As we said in earlier chapters, MEU is silent on what those preferences should be. But it is not silent on how we should reason with probabilities. Prospect Theory adds extra machinery in order to make a theory of decision making that fits with systematic observations of humans. MEU thinks that humans are better than that. And where they are not, there are opportunities for us to train ourselves to become better.

In other words, Prospect Theory, in its capacity to capture human decision making, can be used to both improve our choices, but can also introduce maladies. Maximize expected utility theory, by contrast, can be seen as a way to help us remedy those maladies. The place to start is with training ourselves to think about probabilities correctly. We will learn some of the basics of probability theory soon.

6.6   Summary

The goal of this chapter was to answer three questions. First, what is the scope of MEU? Here we answered that MEU is a theory of what we called standard decision making. We looked at clear cases of standard decision making, clear cases outside of it, and an example (choosing to have a child) that can be described as being in between, which we called an instance on non-standard decision making. These considerations help us understand what sorts of examples of decision making we should be thinking of when we say that MEU is a theory of it.

The second question was about reasons in favor of MEU given its scope. Here we considered a "long run" argument, which says that MEU is a good theory of standard decision making because in the long run MEU will provide higher payoffs than the alternatives, just as we would expect that good standard decision making will produce higher payoffs in the long run.

The third question was about arguments against MEU. We divided these in to two. The first set of objections are theoretical in that they attempt to demonstrate that the internal workings of MEU are not coherent. These objections apply directly to normative versions of the theory, and perhaps to some extent descriptive versions of MEU as well. The second set of objections came from empirical considerations. These objections were primarily focused on MEU as a descriptive theory, but they can also be turned into arguments against the normative version of it as well under careful characterizations.

6.7   Exercises

1. Suppose, contra L.A. Paul, that we *could* set up the decision to have a child as a standard decision (pretend like the experience is not transformative). Using the four relevant grous (Luck Parents, etc.) to build a decision table. What utilities would you assign to each outcome? What probabilities do you think these get? And what are the expected utilities for the two choices?

2. In the movie *The Matrix* Neo must choose between the red pill and the blue pill. Is Neo's choice an instance of standard decision making, or is it more like L.A. Paul's idea of a transformative experience? Explain. (If you haven't seen the movie, think of another example that might count as a transformative experience and present some reasons for thinking why it is one and why it might not be one.)

3. Which objection against normative MEU do you find most com-

pelling? Does the long run argument have a response to that objec-
tion?

4. Which objection against descriptive MEU do you find most com-
    pelling? To what extent are you convinced (or not) that Prospect
    Theory can salvage MEU? Or is Prospect Theory too different that
    you think we should consider it to be a different theory?

7
Intervention

We have emphasized the importance of distinguishing between that which is in our control and that which is not. After all, it seems to be a constitutive feature of making a decision among a number of choices. If we can't control something, it hardly seems correct to think that we have a choice in the matter.

Another way of thinking about what is and isn't in our control is in terms of **intervening** on the world. That is, when we make a decision by selecting an action, we are intervening on the world by having a causal effect on it. When I choose to study, a different set of worldly facts come about than if I choose to party: if I study, my brain will have some information accessible to it that it would not have if I had partied (not to mention the lack of consequences from partying). If I go to the park to play frisbe with my friends, a different set of events will unfold than if I had chosen to stay home to work on a project. If I pick the one box in the Newcomb problem that will bring about a different result than if I had picked the two box option.

We have to be careful about how we think about intervention in our decision tables. It can be tempting to think that by intervening we are shaping the world to our will, but that's not quite right, or at least not that simple. A better way is to imagine that the world is like an agent that has already selected which column in a table is true, but we are uncertain about what the world has selected. The action that we then select intervenes by narrowing down the set of possible outcomes by selecting a particular row. The world then "reveals" what it had already selected just before we made our choice, and the outcome then becomes known to us.

There have been some recent developments in modeling the idea of interventions. These include the idea of a causal model.[1] This chapter will introduce the idea of causal models and their representations as causal graphs (or networks). After introducing the basic ideas behind causal models, we'll use them to illustrate that there

[1] Judea Pearl's book *Causality* had a big impact on thinking about causal models. Another more recent book of his is *The Book of Why*.

are different kinds of decisions theories. All of these make use of ex-
pected utility, but they differ in their interpretations of how choices
intervene.

7.1   Causal Models

To start, let's temporarily assume that the world is perfectly deter-
ministic and that every single cause has a single effect, and every
single effect has a single cause. For example, suppose we have the
very simple example that you studying for a test guarantees that
you pass it, and that passing the test will cause you to win a prize.
Here we are implicitly suggesting that if you don't study then you
won't pass, and that if you don't pass you won't get a prize. The way
we represent this example is by having a *variable* for each cause or
effect, and each variable can have a set of values. For example, our
first variable would be BEHAVIOR and it has two possible values,
study or party. Our second variable would be TEST and have values
of pass or fail. Our third would be PRIZE with values yes or no. We
can then use arrows between these variables to represent their causal
connection:

BEHAVIOR ⟶ TEST ⟶ PRIZE
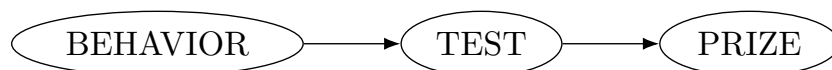
Figure 7.1: Simple causal pathway

Given our assumptions, there is just one way for you to get the
prize. You getting the prize depends on you passing the test, and
you passing the test depends on you studying. We can express these
dependencies using some very simple equations. Let's use '1' to
represent cases of studying, passing, and winning the prize, and
we'll use '0' to represent partying, failing, and not getting a prize. In
addition, we'll read 'X:=Y' as 'the value of X depends on the value
of Y'.[2] The following single equation then represents that the PRIZE
variable depends on the TEST variable:

$$PRIZE := TEST$$

and similarly the TEST variable depends on the BEHAVIOR variable:

$$TEST := BEHAVIOR$$

Given these equations, we can imagine there are three ways to ensure
that PRIZE = 1.[3] The first, as if by the power of God, is to simply
just make PRIZE = 1. This is a bit awkward, for effectively we are
''screening off" all the upstream causes in the graph as if they didn't
really matter after all. The second, which is less arbitrary (given our

[2] Note the direction of dependency,
which only goes the one way. Think
of ':=' as a single symbol, like going
backwards along the arrow on the
diagrams.

[3] Heads up: the first two are weirder
than the third.

other assumptions), is to make TEST = 1. By the equation $PRIZE :=$ $TEST$ this makes PRIZE = 1. But this is still not the way that we're thinking about our example, particularly give the way we have drawn our simple graph. The only way to make PRIZE = 1 given what is in our control is to set BEHAVIOR to be 1. This in turn makes TEST = 1 given that $TEST := BEHAVIOR$ and then PRIZE = 1 given that $PRIZE := TEST$. In the causal modeling literature, all three of these possibilities we've covered would count as a kind of "intervention", but what they have in mind is an intervention on the model. What we mean here by 'intervention' is being in control of a variable by being able to set its value.

There is, nevertheless, a sense in which getting the prize is in your control, even if you can't directly intervene on that variable. When we set $BEHAVIOR$ to 1, that fact, combined with the causal graph, will have guaranteed downstream consequences that ultimately make $PRIZE$ equal to 1.

The simple scenario we have been working with might remind you of decisions made under certainty. Our primary interest, however, has been dealing with decisions where there is some amount of uncertainty, i.e. where there are some things that are outside of our control that combine with our decision to generate an outcome. For example, suppose you are playing a game where your friend is going to flip a fair coin and you can choose to either bet or not bet while it is in the air. If you bet and the coin lands heads, you win a dollar. If you bet and the coin lands tails, you lose a dollar. And if you choose not to bet at all then you don't win or lose anything. In our standard matrix representation, this scenario looks like this.

	Heads	Tails
Bet Heads	$1	-$1
Don't Bet	$0	$0

We can use our graphs to also represent this situation. We'll have three nodes, one for the possible actions (bet / don't bet), one for the possible world states (heads / tails), and one for the possible outcomes (the four cells in our table).

In addition to this graph, we can write down the equation that will determine how much money you would receive as the outcome. As the graph suggests, this depends on both your action and the world, though you only have control over the former. We'll use '1' to represent that you bet and '0' that you don't, and '1' to represent that the coin turned up heads and '-1' that it did not. The equation is:

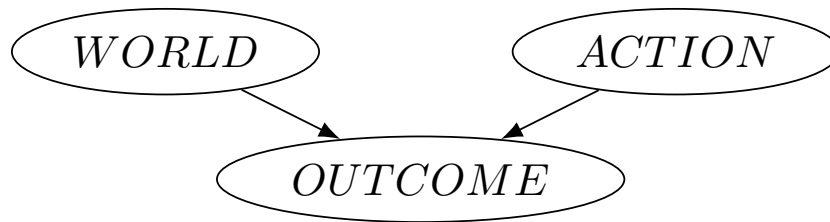$$OUTCOME := ACTION \times WORLD$$

Inserting the relevant combinations of values for the actions and world states will output the corresponding values in the outcomes.[4]

We now have three kinds of representations we can make use of.[5] The graph representation will be particularly helpful for illustrating some important points.[6] In general, a causal graph is a collection of nodes with arrows between them. An arrow represents that a node (a variable) has a direct influence on another node (the one that the arrow is pointing to). The node where the arrow starts is called the parent node, and the arrow where the arrow points is the child node. When there is a chain of arrows, a node can have 'upstream' ancestors (parents, parents of parents, parents of parents of parents, etc.), as well as 'downstream' descendants. Since the arrows are representing causation, and causation only flows in one direction, the collection of arrows cannot have cycles, i.e. it is never the case that a node can be connected back to itself by following a chain of upstream arrows (nor similarly downstream ones).

[4] Exercise: What do those values need to be?

[5] That is, tables, graphs, and some equations.

[6] Don't be fooled by the simplicity of the diagrams. They represent a lot of information.

7.2   Common Causes

An important concept that causal graphs help us illustrate is the idea of a common cause. Let's suppose that you occassionally suffer from severe headaches. A pretty good indicator of whether you will get one later in the day is if you have "spots" in your vision in the morning. You also notice that on days when you eat a banana in the morning you tend not to get a headache later in the day. Let's use $H = 1$ to represent that you get a headache, $H = 0$ that you don't, $V = 1$ that you have spots in your vision, $V = 0$ that you don't, and $A = 1$ that you ate a banana in the morning and $A = 0$ that you didn't. The causal graph would look like this.

The equation for the value of the headache variable would be:

$$H := V \times (1 - A)$$

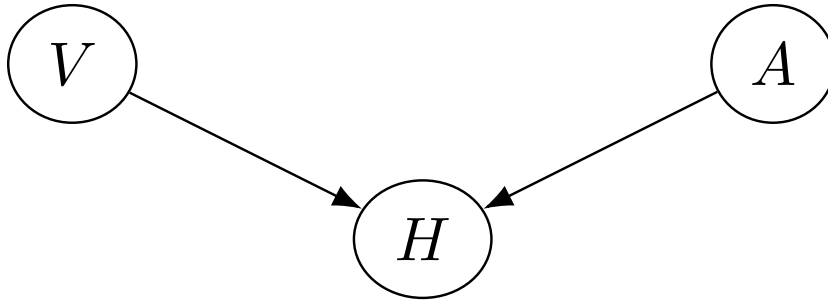It is helpful to see the four relevant possibilities in the form of a table.

Figure 7.3: Getting a headache can depend on having spots in your vision or whether you ate a banana.

V	A	$H := V \times (1 - A)$
1	1	0
1	0	1
0	1	0
0	0	0

When we do some reasoning with these representations, we'll notice that the only time you get a headache is when you have spotty vision but choose not to eat a banana.

Now lets add in an additional consideration. Suppose that you go to the doctor and they note that you are prone to a potassium deficiency. A potassium deficiency, they tell you, can often lead to spots in vision, as well as cravings for bananas (because they are high in potassium!). Suddenly you realize that the days that you had headaches were days when you had both spotty vision in the morning and also had cravings for a banana - but for one reason or another you decided not to eat one. Equipped with this new knowledge, you now make sure you always have bananas on hand so that you can choose to eat one when you have a craving.

Let's update our causal graph. You'll notice that a potassium deficiency has two effects, one on spotty vision and one on your choice to eat a banana. Let's say $D = 1$ means you have a potassium deficiency and $D = 0$ means you don't.
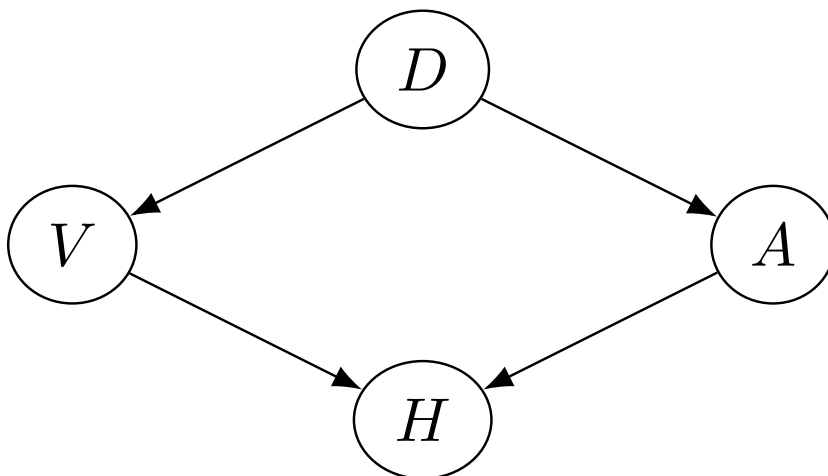
Figure 7.4: A common cause pathway. A deficiency of potassium causes both spots in vision and a desire to eat bananas.

This more complicated causal graph can no longer be represented by a single equation. Rather, we now need a *system* of equations that

$$V := D$$
$$A := D$$
$$H := V \times (1 - A)$$

If we now consider the possibilities, we'll notice that you never get a headache. This is because on the days that you would get a headache from spotty vision are days when you have a potassium deficiency, but a potassium deficiency also means that you will eat a banana, and thereby counteract the headache. In our example, spotty vision and eating bananas have a **common cause**, namely a potassium deficiency.[8]

Causal graphs make it really easy to spot when two or more variables have a common cause. We simply take those nodes and trace the upstream arrows: if the variables all have a common ancestor, then they have a common cause (the variable that is the common ancestor).

When two or more variables have a common cause, those variables will be highly correlated with each other, even though they are not causes of one another. For example, an increase in ice cream sales can be correlated with an increase in pool drownings. We should not think, however, that ice cream sales cause pool drownings (nor for that matter, that pool drownings cause more ice cream sales!). Rather, an increase in ice cream sales and pool drownings have a common cause: an increase in temperature, which tends to cause people to both buy more ice cream and to go swimming (which in some cases unfortunately leads to some people drowning).[9]

Similarly in our case: your spotty vision does not cause you to choose to eat bananas, nor does eating bananas cause your spotty vision. Rather, it is a potassium deficiency that causes both.

7.3   Application to Newcomb-like Problems

Causal models help us gain insight into some of the decision problems we've seen. Recall Newcomb's problem, where a game show host consulted an AI that has profiled you to determine whether to put a million dollars into a second box or not. The payoff table looked like this:

	AI predicts two box	AI predicts one box
One box (just A)	$0	$1,000,000
Two box (A and B)	$1,000	$1,001,000

What this table hides are causal assumptions about the outcomes.

[8] Try out this reasoning on the system of equations: set $D = 0$ and find the value of H, then do it again when you set $D = 1$. What do you notice about the value of H in each case?

[9] Careful here. If X is a common cause of Y and Z, then Y and Z will be correlated. But just because U and W are correlated does not mean they have a common cause - some correlations are *spurious*. See these examples of spurious correlations.

We have been insisting that when we set up decision matrices that the rows are the choices or actions available to us that are in our control, and the columns are the world states - the things that we are uncertain about and are out of our control. Using our simple causal pathway diagram to illustrate this, we might think that the following represents the above table:
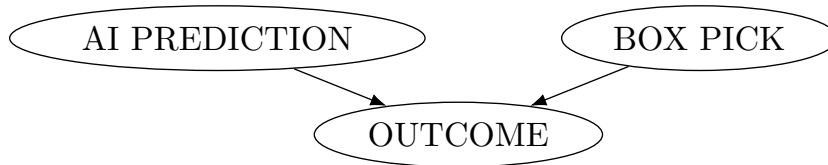

Figure 7.5: Misleading representation of Newcomb problem

But given the narrative of the scenario, there is a correlation between what the AI predicts and what people choose to do. The reason for this correlation is not an accident - after all it is stipulated that the predictions of the AI are highly reliable (even if it's not perfect). When the AI makes a prediction on what someone will choose, it is consulting some body of information that we reasonably could put under the umbrella of a person's *character*. That is, a person's character is a guide to what action they are likely to select (their box pick: one or two) as well as what the AI will predict (which leads to world state: $1M in box A or not). So a better respresenation of this situation would be the following causal graph:
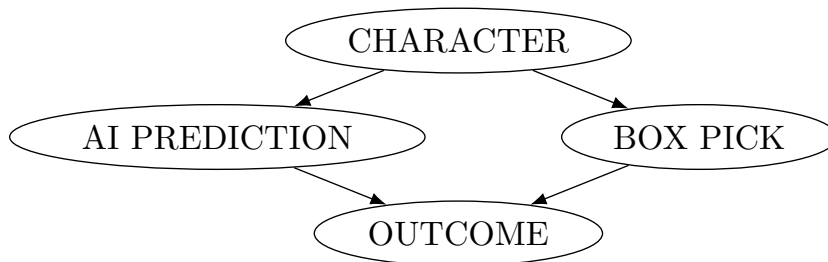
Figure 7.6: Newcomb pathway

This representation makes it clear that there is a common cause to the AI's prediction and your action. If that's right, then the setup is actually not a legitimate kind of decision making problem, since we are not maintaining that the world states and actions be independent of one another. It is not obvious, however, how we might go about fixing this and numerous proposals exist, some of which we'll cover in due course.

First, let's see how some other examples we've discussed are like the Newcomb problem. Consider again the situation where you are deciding whether to work on a project from home or to show up to
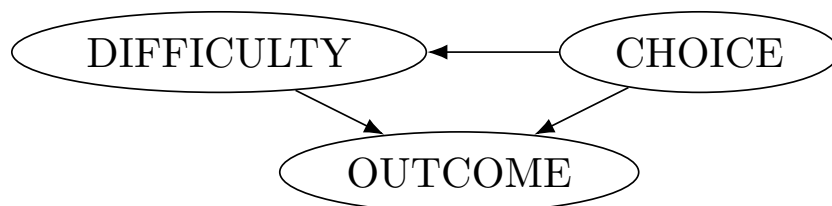
the park to play a game of frisbee with your friends. In that decision problem, the uncertainty involved whether "Annoying Guy" would also show up. Here was the utility table we had.

	Annoying guy stays home	Annoying guy shows up
You stay home	2	0
You show up	3	1

We additionally suggested that, while you might not like it, you and Annoying Guy are very much alike. In fact, let's just suppose that you and Annoying Guy are identical twins. What this means for this example is that what you choose to do is correlated with what Annoying Guy does. We can again use a causal graph to model this situation. When we do, we need to decide whether the correlation between your choice and Annoying Guy's is completely accidental, or whether there is a reason for that correlation. We have suggested that it's the latter. If we unpack this, it would be again reasonable to think that the correlation reflects a common cause between the two of you: you have similar characters.[10] So we might have this kind of causal graph:

[10] And given that we said you are identical twins, the common character could be the effect of having a common physiology, but we leave that out for now.

Figure 7.7: The Twin Example

CHARACTER

TWINS ACT   YOUR ACT

OUTCOME
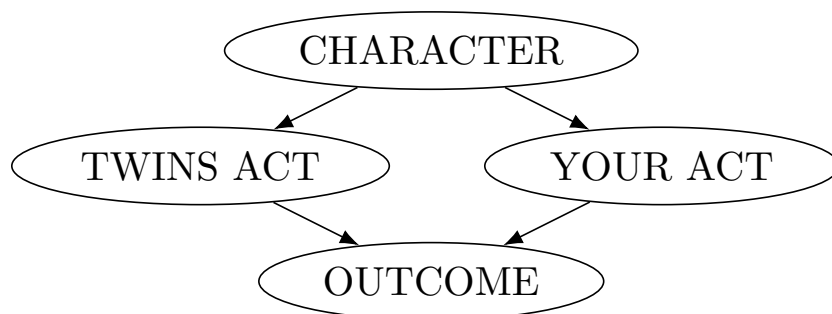
Think about another example we've discussed about whether to study for a test or to party. Here too it might be tempting to set up a graph that looks like this.

DIFFICULTY ← CHOICE

OUTCOME

Figure 7.8: Test Example - Incorrect decision model setup because world states and acts are supposed to be independent, but they are not (CHOICE is a common cause).

But that graph is incorrect. We insisted that we had to be careful to

understand that what it means for a test to be difficult is independent of the action to study - "difficult" had to be some kind of measure by which an instructor uses to design questions. This is because your studying for the test can affect the *perceived difficulty* of it, and that would violate our requirement that the world states be independent of the actions. A causal graph makes it clear why we have to be careful: if we're not, then your choice (study or party) would be the common cause of the outcome!

If world states and our choices are independent prior to the outcome, as we are trying to be careful to do, then we should have this:
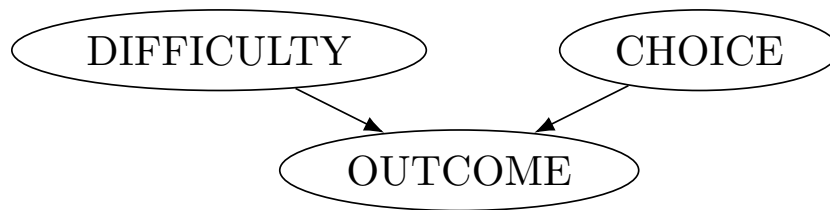
Figure 7.9: Test Example - Correct decision model setup if we insist that world states and acts are supposed to be independent.

What we might notice is that the Newcomb problem is more like the incorrect test example if we take into account something like a person's character. Newcomb-like problems all seem to have a kind of causal structure where, if we go back far enough, there are common causes that influence both the world state and our action. That is, if we were to map out causal graphs in more detail, they are likely to end up looking something like the following.
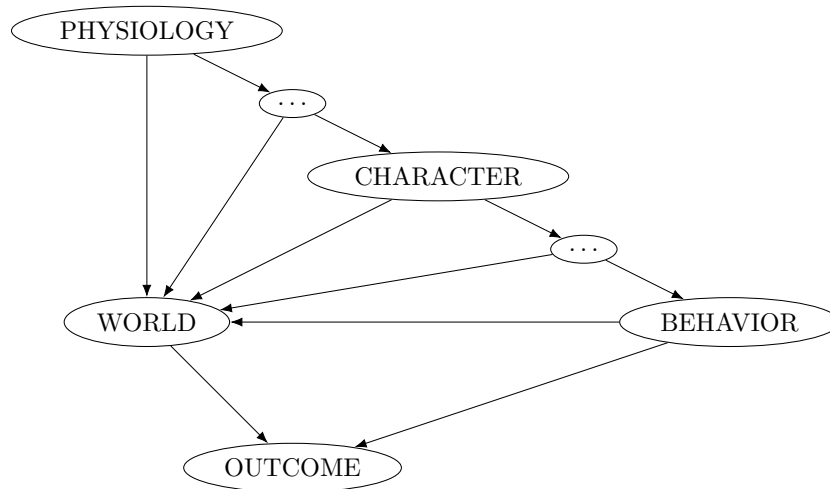

Figure 7.10: More realistic causal graphs of decisions. The idea to represent things in this way come from Kenny Easwaran's 'A classification of Newcomb problems and decision theories' in *Synthese* (2021, 198).

What are we to make of these kinds of situations? Isn't it true that our choices can impact several different variables in this chain? Regularly choosing to exercise can impact our physiology - that's

the whole point behind training! And in the banana example it can ensure that we don't get a potassium deficiency, and if we do we still can prevent our headaches. And isn't it the case that the way we develop our character is by regularly making the kind of choices that correspond to the kind of person we want to be? In the Newcomb problem, we want to be the kind of person that picks the one box, since that's the kind of information the AI would use to make its prediction, which in turn makes it highly likely that there will be a $1M in the box! Do these kinds of considerations mean the world state is in a person's control after all?

We have been using the idea of control in a binary (on/off) kind of way, but as the causal graph above suggests, what is and isn't in our control might not be such a clear thing. Perhaps a better way to think of control is that it comes in degrees. Towards the left side of the causal graph are things we have less control over, like our physiology, and towards the right are things that we have more control over, like our behavior in a particular moment.

Whether control is binary or comes in degrees, what we are drawing attention to is a question about where the locus of a choice is. We will see shortly that how one answers this question will largely determine the type of decision theory one ends up endorsing.

7.4   *The Locus of Choice and Types of Decision Theories*

It is reasonable to suggest that rationality has a kind of unity. We started our analysis of decision making by defining instrumental rationality - an action is more or less rational depending on how well it helps achieve some goal or aim. Similarly, there is a principle that connects what is possible for us to do and what we ought to do: if something is not in our control to do, then it's false to say that we have an obligation to do it. One way to wrap up some of these considerations is with a kind of "enkrasia" principle along the following lines:

(Unity of Rationality) It is rational to do X if and only if: i) it is rational to *plan* to do X, and ii) it is rational to have character traits that lead to planning to do X, and iii) it is rational to have the kind of physiology that allows for the development of such character traits, and so on.

If the unity of rationality is true and we are trying to maximize expected utility, then for Newcomb-like problems we should be selecting the analogs of the one-box option. For if we did not, then selecting the two-box option in the moment is a reflection of our broader character, which means the AI would have made that prediction, and so we would be missing out on $1M (and consistently so if

we went on the game show multiple times).

The dominance argument suggests, however, that the rational thing to do is to pick the two-box option. How can this be rational if that effectively guarantees that one will be consistently missing out on $1M? There are at least two things to say here. First, one might suggest that the world is imperfect and sometimes it rewards what is irrational. Just because the world rewards the irrational does not detract from the rationality of the choice.[11]

The second thing one could say is that rationality might not have the sort of unity that is suggested above. That is, perhaps in some situations it is possible that the rationality of doing X is at odds with the rationality of planning to do X, or is at odds with the rationality of having the relevant character traits. That is, if we deny the unity of rationality, one might say that it is perfectly consistent to develop a character that corresponds to being a one-box kind of person, but then in the moment on the game show where one makes the choice, one chooses to two-box at the very last minute. The dominance argument seems to be suggesting something along these lines if the outcome of winning $1M + $1,000 is a live possibility at all.

A related difficulty concerns about our *knowledge* of what is and is not in our control, as well as our *knowledge* about the kind of character or physiology we have. In fact, it's not unreasonable to think in some cases our brains are planning to do something what we aren't consciously aware of (yet), and so in a sense we might not have knowledge of our own plans. How might that affect our considerations about the unity of rationality? Let's illustrate with an example.

There is a parasite called *toxoplasma gondii* that is primarily hosted by cats but affects mammals in general. Infected individuals tend not to show symptoms, but there do exist a wide range of negative effects. Interestingly, infected rats will tend to be less afraid of cats and even be attracted to cat urine, which from the parasite's perspective is advantageous because that's how it perpetuates its life-cycle (the larvae develop in rats and then mature in cats, laying more eggs in the excrement of cats which are then picked up by rats).

Consider then an agent that knows it's possible they have toxoplasmosis. Given this possibility, they know that there is some chance that they will experience negative effects from it, but they also know that they are more inclined to adore cats by, e.g., speaking to them, collecting pictures of them, and potentially even petting them. That is, they consider the following causal graph.

Let's assume that while the chances of negative side effects are small, the negative utilities of those side effects are very large, whereas the utilities that come with adoring cats is relatively small in comparison. Should the agent decide to adore cats or not?

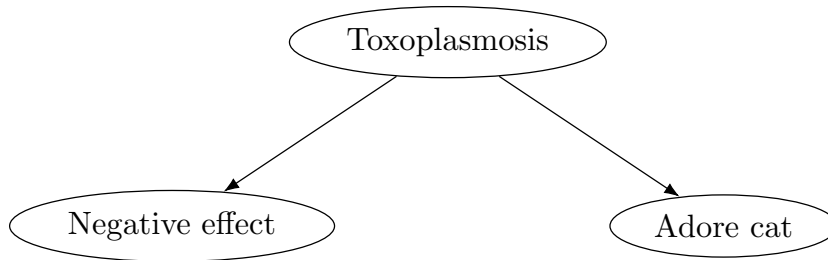[11] David Lewis has made this sort of response in his 'Causal Decision Theory' in 1981.

Toxoplasmosis

Negative effect   Adore cat

If the agent uses the causal graph as is, they might reason that choosing to adore cats is itself evidence of having toxoplasmosis, which in turn makes it more likely that they will have negative effects. Note that their choice *does not cause* the negative effects. Rather, when they consider the action of adoring cats, they in turn use that as new information about themselves! And that in turn might influence what they end up deciding to do. We don't yet have the machinery from probability theory to make this more precise, but know that so long as the negative utilities are sufficiently larger than the positive utilities, and the probabilities for negative effects aren't too small, the agent will choose not to adore cats.

Alternatively, the agent might follow our insistence that there be no common causes between the world states and our relevant choices. So to the extent that they have a choice in adoring cats or not, this will be independent of the already settled fact of whether they have toxoplasmosis or not. To represent this, they modify the causal graph so that they severe the dependence arrow from toxoplasmosis to adoring cats:
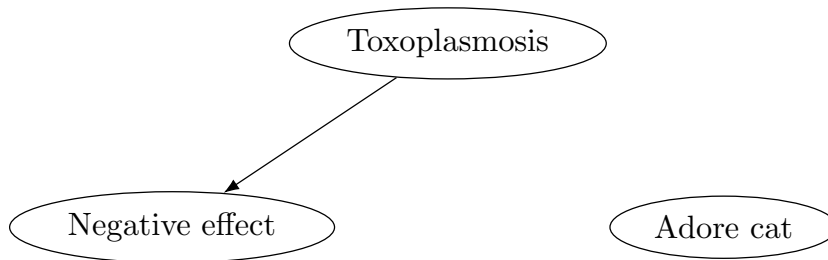
Toxoplasmosis

Negative effect   Adore cat

In this way of reasoning, the agent recognizes that adoring cats is evidence that they have toxoplasmosis, but this is already accounted for in the other two nodes and the connection between them. In this reasoning, toxoplosmosis does not cause the *choice* of adoring cats. So now when they reason about what to do, they recognize that adoring cats always adds some small extra utility (say 1), regardless

of whether the world state is that they have toxoplasmosis or not (where toxoplasmosis has a negative utility of 1000, say).

	Has toxoplasmosis	Does not
Ignore cats	-1000	0
Adore cats	-1000 + 1	0 + 1

From this perspective, the agent reasons that they should adore cats, since that's the dominant option.

What we are illustrating with this last example is a question about whether choices, wherever they are located in the causal graphs, can be informative about the very causal graphs we are using to guide our choices.

So we have two kinds of questions. One kind of question is about

WHERE THE LOCUS OF CHOICE IS. The other kind of question is about

WHETHER INTERVENTIONS RELATED TO CHOICES ARE FURTHER EVIDENCE to be used in our reasoning. It turns out that how we answer these questions will characterize different kinds of decision theories.

The most common decision theory, called

CAUSAL DECISION THEORY (CDT), puts the locus of choice furthest to the right of the causal graph we had above, i.e., a choice for the purpose of standard decision making is what decides the behavior. Variables that are upstream, such as character and physiology, are not part of what is being analyzed, at least not for the purpose of understanding rationality in terms of comparison of options.

On the other extreme of the causal graph, and a more recent addition to the literature, is

FUNCTIONAL DECISION THEORY (earlier known as **timeless decision theory**). On this decision theory options are compared by considering the furthest upstream points of evaluation. So if that happens to be character, for example, then options are compared by imagining the impacts of interventions (the removal of causal arrows) at those points. In some presentations of FDT, we imagine that we are choosing what kind of robot or set of functions we do best in a decision problem.

A third kind is

EVIDENTIAL DECISION THEORY (EDT). The idea behind EDT is that
we don't intervene on a causal graph at all when we analyze what to
do. Rather, just like in the toxoplasmosis example, we simply condi-
tionalize on the evidence that is captured on the graph. This notion
of conditionalization has a technical meaning that is addressed in
probability theory. So it is finally time for us to get clear on what we
mean when we talk about probabilities.

Exercises

1. Consider the following system of structured equations:

$$O := W \times A$$
$$A := P$$
$$W := C$$
$$P := C$$

   Draw the corresponding causal graph.

2. Here is an example to consider from Derek Parfit (1984): Suppose
   that I am driving at midnight through some desert. My car breaks
   down. You are a stranger and the only other driver near. I manage
   to stop you, and I offer you a great reward if you rescue me. I
   cannot reward you now, but I promise to do so when we reach
   my home. Suppose next that I am transparent, unable to deceive
   others. I cannot lie convincingly. Either a blush, or my tone of
   voice, always gives me away. I want to get the major benefit of a
   ride out of the desert, but actually giving you the reward is a cost
   to me. So ideally, I would like to be able to formulate a sincere
   plan to give you the reward, a plan that is sufficiently sincere that
   you come to believe me and give me a ride out of the desert, but
   after having gotten the ride I'd rather not follow through on this
   plan.

   • Using the causal graph model representation, analyze this deci-
     sion (note that planning happens somewhere between CHAR-
     ACTER and BEHAVIOR). You may find it helpful to also in-
     clude the decision table like we have been doing.

3. Here is an example from Egan (2007): Paul is debating whether
   to press the "kill all psychopaths" button. It would, he thinks,
   be much better to line in a world with no psychopaths. Unfortu-
   nately, Paul is quite confident that only a psychopath would press
   such a button. Paul very strongly prefers living in a world *with*
   psychopaths to dying.

- Represent Paul's decision problem in a standard decision matrix. Then draw a causal graph diagram for it. Is there a common cause here? This is sometimes presented as a problem for causal decision theory. Explain how that could be.

4. In our discussion of the unity of rationality, we considered the possibility that an action is rational even though "the world" would reward the irrational action (at least according to dominance reasoning). Can you think of any examples besides our discussion of the Newcomb Problem or the Toxoplasmosis Problem? What argument would support the claim that the example action is rational despite the world rewarding a different action (is it the dominance strategy or something else)?

8
Odds, Probabilities and Actions

We have come to the point where it will no longer be enough to work with our intuitive notion of probability. If we want to go beyond toy decision problems, like the one's we've been covered, and model decisions that are more like the ones we more frequently face, we'll need to have a sufficiently robust understanding of probability. For example, one of the main things we'll want to be able to do is update our beliefs given new information, which effectively amounts to knowing how to update probabilities.

Probabilities are a way of quantifying beliefs. It may seem impossible to measure something as elusive and subjective as beliefs. But some clever conceptual tools have been developed to do just that.

The key idea behind measuring a person's belief about the world is to figure out how willing they are to risk things that they care about. To illustrate this idea we're going to momentarily assume that money is our measure of utility.

8.1   Odds and Fair Betting Rates

Here's a roughly general observation of people's behaviour: the more confident someone is that some event is going to happen, the more willing they are to bet.

Suppose *S* is some event that Bob and Ally care about. The event might be a sports team winning a game, that it's going to rain tomorrow, that the stock price of some company will be higher by next year, etc.

Let's say Bob is more than 50% confident that *S* will happen. In fact, let's suppose that Bob would accept a deal that would pay him $1 if *S* happens, but would cost him $2 if it doesn't. Let's say Ally thinks Bob is wrong to be so confident and agrees to take Bob's bet. In effect, this means that Ally is willing to put $1 on the table for the chance of winning the $2 that Bob is willing to put on the table. The

stake is the sum of all the money on the table, in this case $\$1 + \$2 = \$3$. Whoever ends up being right gets to take the stake.[1]

BOB'S FAIR BETTING RATE can be expressed by dividing his potential loss by the stake:

$$
\begin{aligned}
\text{Betting Rate} &= \frac{\text{Potential Loss}}{\text{Stake}} \\
&= \frac{\$2}{\$2 + \$1} \\
&= \frac{2}{3}.
\end{aligned}
$$

Bob's betting rate is 2/3, which is a reflection of how confident he is that S will happen. And here's the next move: that's Bob's personal probability that S will happen, i.e. $Pr_{Bob}(S) = 2/3$.

BOB'S FAIR ODDS is another way that betting is sometimes talked about. To express Bob's fair better rate in terms of odds, we take the ratio of potential loss to potential win:

$$
\begin{aligned}
\text{Odds} &= \text{Potential Loss} : \text{Potential Win} \\
&= 2 : 1
\end{aligned}
$$

The odds that Bob would accept are another reflection of his degree of confidence. In fact, there is a handy way of linking up our notion of expected value with odds and probabilities.

A FAIR BET is one in which the *expected value* is zero. That is, if we weight the potential win by the probability of winning, and we weight the potential loss by the probability of losing, the odds should "cancel out" or "wash out":

$$(2/3)(\$1) + (1/3)(-\$2) = 0.$$

Here's a helpful visual way of understanding the idea of a fair bet. Notice first an inverse relationship between probabilities and payoffs when it comes to risks (especially in gambling): events with really high payoffs tend to have low probabilities, and likewise, the more probable an event is the lower the payoffs tend to be. If the amount of probability is like the width of a rectangle, and the payoff (or loss) is like the height of a rectangle, then a fair bet will be one in which the area of a rectangle that represents Bob winning will have the same amount of area that represents Bob lossing.

To be clear, what Bob considers to be a fair bet might change. For example, suppose Bob comes into some information that significantly decreases his confidence that S will happen. Let's say his confidence

goes all the way down to 10% (i.e. 1/10). Since his confidence went down, he should be willing to risk less, that is, he should be willing to stake much less. How much less? As a fair bet, Bob will want to make sure that the expected value will be 0:

$$(1/10)(\$9) + (9/10)(-\$1) = 0.$$

So for Bob to be willing make a bet with Ally give this new information, Ally would need to be willing to put at least $9 in the stake for Bob's $1.

Notice how our visualization using rectangles will change for this new scenario, but the two rectangles will still have the same area.

HERE'S A GENERAL RECIPE for quantifying a person's probability that a proposition S is true using the idea of fair bets:

1. Find a bet on S that they see as fair. Call the potential winnings W and the potential losses L.
2. Because the bet is fair to their eyes, set the expected value of the bet equal to zero:
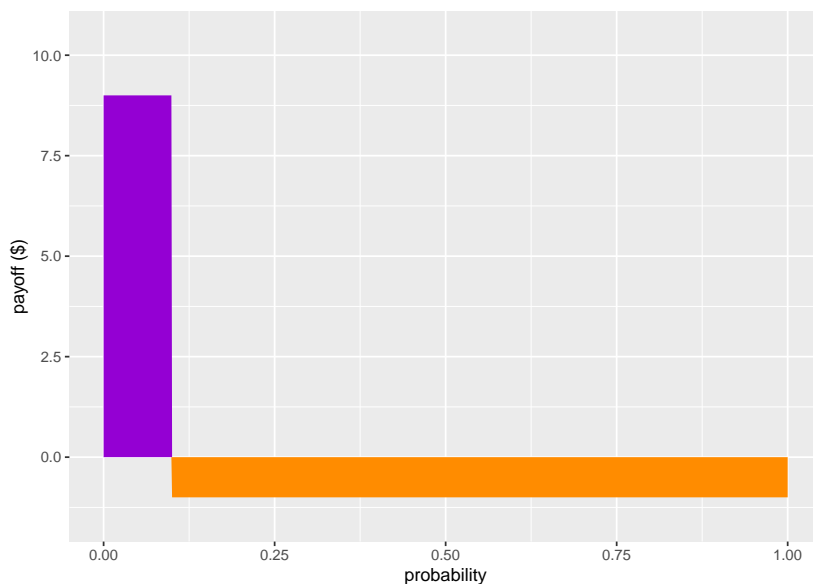
$$[Pr(S) \times W] + [(1 - \Pr(S)) \times -L] = 0.$$

Figure 8.2: A bet that pays $9 if Bob wins and costs $1 if he loses is fair when the probability of winning is 1/10.

3. Now solve for $Pr(S)$:

$$(Pr(S) \times W) + ((1 - Pr(S)) \times -L) = 0$$
$$(Pr(S) \times W) + (-L) + (Pr(S) \times L) = 0$$
$$(Pr(S) \times W) + (Pr(S) \times L) = L$$
$$Pr(S) \times (W + L) = L$$
$$Pr(S) = \frac{L}{W + L}.$$

Notice that we have the formula for the fair betting rate again! It's helpful to memorize this formula so you don't have to do the derivation each time. But more important than that is knowing that there is a recipe for getting from bets to personal probabilities.

8.2   Advantageous Bets

The general recipe we just developed uses the idea of fair bets (or fair odds). The idea here is that a person would be willing to take either side of the bet. Using the language of preferences: the person is *indifferent* between the two options. What happens if betting rates aren't fair in the eyes of Bob? In that case Bob would no longer be indifferent, he would want to take one side!

For example, let's suppose that Bob is confident that S will happen, say by three to two odds, i.e. $Pr(S) = 3/5$. That means that, from Bob's perspective, for every $2 that Ally is willing to put in the stake, he is willing to put in $3. That is his fair bet.[2] Now consider

[2] Notice: $(3/5)(\$2) + (2/5)(-\$3) = 0$

two scenarios:

- (Scenario 1) Ally is willing to give Bob even odds, i.e., she is willing to put in $3 for every $3 that Bob puts in the stake (and as before, if S happens then Bob wins the stake). So for the same amount that Bob is willing to lose in *his* fair bet, in this scenario he has the same probability of winning, but now the *payoff* for winning is higher. From Bob's perspective, his expected value is a gain:

$$(3/5)(\$3) + (2/5)(-\$3) = 0.6$$

So Ally's offer would favor Bob and is advantageous (in his eyes).

- (Scenario 2) Ally is only willing to put in $1 for every $3 that Bob puts in the stake (and as before, if S happens then Bob wins the stake). So in order to win $2 from Ally in this scenario, he would have to be willing to lose $6, which is twice the amount relative to his fair better rate! From Bob's perspective, his expected value is now a loss:

$$(3/5)(\$1) + (2/5)(-\$3) = -0.6$$

So here Ally's offer would not be advantageous to Bob.

THERE'S AN IMPORTANT LESSON HERE. Just because a person is willing to take a bet does not mean that they think it's a fair bet. Like in Scenario 1, Bob is willing to take bets that he perceives as advantageous to him. But in Scenario 2, he would rather be in Ally's shoes! So when we're measuring people's personal probabilities, we have to make sure that the bets that they are willing to take are ones that they think are fair (i.e. the bet, from their perspective, isn't advantageous or disadvantageous).

THERE'S COMPLICATIONS THOUGH. Recall that money doesn't perfectly track utility, it is at best a very rough estimate. In previous chapters we saw how gaining a dollar is not the same as losing a dollar. Moreover, the utility of gaining/losing a dollar depends on how much you already have.

ANOTHER COMPLICATION is that we're assuming that Bob is *following* the expected utility strategy. As we saw in the discussion of paradoxes like the Allais paradox, we have some reason for thinking that people don't follow it. So our way of measuring personal probability isn't perfect, but it's still a significant advancement in measuring something that we initially thought was purely subjective.

BUT HOLD ON. It's not a defect of a camera when it won't take pictures in a dark room. Similarly, if someone isn't following the expected utility formula, that says something about the person, not the method by which we can measure personal probabilities.

OUGHT PEOPLE TO BEHAVE according to the expected utility formula? According to the laws of probability and some assumptions that connect beliefs with actions: yes we should. The arguments here are known as *Dutch Book* arguments. Before we get to these arguments, we need to briefly say at least informally what the laws (or axioms) of probability are.

8.3   *The Axioms of Probability and Dutchbooks*

There are three standard axioms of probabilities. I'll present them informally here.

1. (Non-negativity) Probabilities can't be less than zero, i.e., $P(S) \geq 0$.

2. (Normality) If we sum up the probabilities of *all* possible events, we get 1. For example, if *all* the possible events for tomorrow are Rainy, Cloudy, and Sunny, then the probability that tomorrow is Rainy or Cloudy or Sunny is 1. Another example: if *all* the possible winners in the race are numbered 1 through 7, then the probability that one of them wins is 1.

3. (Finite Additivity) For mutually exclusive events, their disjunction is additive. For example, suppose there's a horse race with four horses: Gumption, Gallifray, Tungsten, and Shadow. If one horse wins, all the others lose (i.e., horse winnings are exclusive events). So, if the probability of Shadow winning is 0.1, and the probably of Tungsten winning is 0.2, then the probability of Shadow or Tungsten winning is 0.3.

When we identify axioms of a theory, we can use those axioms to define other constraints or features of the theory. For example, we can define the probability for the negation of a proposition (i.e. that an event will not occur), $\neg X$ as:

$$P(\neg X) = 1 - P(X)$$

As we've seen above, that means that fixing the probability of a proposition automatically fixes the probability of its negation (and vice versa). For example, if $P(X) = 0.7$ then $P(\neg X) = 0.3$.

This is a helpful feature. Suppose, for example, that we have a lottery with 100 tickets. Each ticket is picked out by a proposition, $X_1, X_2, \ldots, X_{100}$. The probability of wining a given ticket, X_i is

$$P(X_i) = \frac{1}{100} = 0.01$$

Since we have fixed the probability that X_i is true, this automatically fixes the probability that it is false, or equivalently, that $\neg X_i$ is true. Formally,

$$P(\neg X_i) = 1 - 0.01 = 0.99$$

This next claim builds on some logic that we will get to shortly, but it's worth making explicit now as it will play a central role in the argument in the next section. Consider the exhaustive disjunction of all mutually exclusive propositions (e.g., we think about ALL the lottery tickets, and the rule is that when you win with ticket X_i that means the other tickets did not win). This exhaustive and mutually exclusive disjunction will be a tautology. In our lottery example:

$$P(X_1 \vee X_2 \vee \ldots \vee X_{100}) = 1$$

In short, even if there is more than one event that we're considering, the probabilities of all events have to sum to 1.[3]

[3] We'are assuming that the events are mutually exclusive and exhaustive.

THE DUTCHBOOK ARGUMENT says that you should accept the above three axioms of probability because if you don't then you allow for the possibility of guaranteed loses of "fair bets'' - bets that rationality suggests would not be fair at all, but you would be unable to provide reasons for not accepting them.[4]

The Dutchbook argument can be stated formally, but it's enough for our purposes to illustrate the argument using the popular example of horse racing. The following table lists all the names of four horses in a race, including the odds and the implied probability of each horse winning. In addition, the table lists the cost of a bet and the corresponding payout if the horse wins.

[4] The person that does accept the three axioms of probability can say why the bets are not fair because they violate one or more of the axioms.

Horse	Odds	Implied.Prob	Bet	Payout
Gumption	1:1	0.5	$100	$100
Gallifray	3:1	0.25	$50	$150
Tungsten	4:1	0.2	$40	$160
Shadow	9:1	0.1	$20	$180

TOTALS 1.05 $210 $200

Here's what to notice from this table. Suppose you're the bookie and you get one person to bet for each horse. No matter which horse

ends up winning, you as the bookie will pay out $200 from the stake (the sum of the bets you received). But if we look at the total that makes up the stake, that total is $210. The means as the bookie you are *guaranteed* to make $10 on the race.

In point of fact, this is how real races are typically organized - and this is a way for you to calculate how much is being "skimmed off the top". For the most part this isn't really an issue because we're willing to let some amount of money to go to those who are organizing the event. But let's put aside those types of considerations and think instead of a set of bets that you and your friends would set up that is regarded as fair, i.e., that there is no way for someone to scheme and guarantee that they win some money. Each of you in the group has the chance of losing or winning and no one is "skimming off the top".

Looking at the table again with this in mind, which of the probability axioms is being violated? Recall that the probability of the exhaustive and mutually exclusive disjunction of the options should sum to 1, but the total in the above table is 1.05. This is a violation of Normality.

A VIOLATION OF NORMALITY where the sum of probabilities is larger than 1 is good for a bookie because it guarantees they earn money.

BUT A VIOLATION OF NORMALITY where the sum of probabilities is smaller than 1 is bad for the bookie, but good for the gamblers! Suppose, for example, that Shadow withdraws from the race. Now the implied probabilities sum up to 0.95. If now a gambler bets on all the remaining horses (but saves $20 because they don't bet on Shadow), the gambler is guaranteed to make a profit of $10. As you can imagine, this rarely happens in real life, as bookies are quick to adjust their betting offers to reflect these sorts of changes.

The point of the Dutchbook argument is to get you to appreciate that if you think about probabilities in a way that does not respect the three axioms, you are at risk of being put in a position where you are guaranteed to lose. Of course lots of decisions under uncertainty have the possibility of losing - that's simply the nature of risk. But those decisions also typically offer the possibilities of gains - that's why we might be interested in such decisions in the first place. What the Dutchbook argument says is that if you violate any of the three axioms, then included in the decisions you should be willing to accept are decisions where there is a guaranteed loss. It is thought that no coherent account of rationality should find this acceptable. Moreover, such situations are easily avoided by someone who accepts the

probability axioms, because the axioms provide them with a reason to reject such decision offers.

If we're convinced by the Dutchbook argument, which means we agree that we should be striving to think about probabilities as prescribed by the axioms, then this puts serious pressure on the idea that Prospect Theory is a viable approach to capturing a more "human" kind of rationality. Recall that in response to weighting effects, where people treat outcomes with low probabilities differently than outcomes with high probabilities, Prospect Theory proposed a different function to capture how people reason across these contexts. Such a function explicitly deviates from probability theory. Perhaps as a purely descriptive theory that attempts to capture the systematic ways that people go wrong, this is fine. But it would be misleading to think of that account as closer to "human" rationality. After all, we have the capacity to acknowledge the Dutchbook argument. To this extent, at least when we are sober and closer to the better versions of ourselves, we have the capacity to reason according to probability theory.

8.4   Application

Examples from gambling help establish connections between probabilities and our levels of confidence through the idea of odds. In many real world applications, however, the probabilities of outcomes are not as neat as they are in the case of gambling. Nevertheless, it is possible to get some approximations; that is a large part of what the field of statistics is about. We will not go into such details here. But it is worth showing how we can connect some of the concepts so far, and motivate by example what will spell out more abstractly in the next chapter.

Let's use an example where we estimate some probabilities from frequency data, something that has important ties with medicine and epidemiology. In the context of health it is common to represent data in the form of what is called a *two by two table* (or 2x2 table). The idea is to use the rows to represent "the exposed" options, like smoking or not smoking, or like using a hand cream or not using a hand cream. The columns represent "the diseased" and "not diseased" which in our case is "rash continues" and "rash gets better".[5]

Here's an example of what such data might look like. We're supposing that each cell represents the number of participants in a study that met the pair of conditions. For example, there are 223 subjects that used the hand cream and had the rash get better. There are 107 subjects that didn't use the hand cream, but the rash got better anyway. In total there are 330 subjects who's rash got better. The cells

[5] In the smoking scenario the columns could be "hypertension'' and"without hypertension'' for example.

under "Rash Continues" is read similarly. Notice that the cell in the most bottom right corner is the total number of subjects in this study: 426.

	Rash Gets Better (R)	Rash Continues (nR)	Total
Used Cream (C)	223	75	298
Didn't Use (nC)	107	21	128
Total	330	96	426

Suppose we asked what the probability is that someone in our sample used the hand cream.[6] To calculate that, we would take the total number of people that use the hand cream (298) and divide that number by the total number of subjects (426). This gives us $298/426 = 0.6995$ (rounded to four decimal places). Similarly, we can calculate the probability that a subject had their rash continue by taking that number (96) and dividing it by the total number of subjects (426). This gives us $96/426 = 0.2254$. The other two probabilities are calculated similarly: take the total from the row or column, and divide by the total number of subjects (426).

The strategy we are deploying in the case of two by two tables has its equivalent in the strategies we are about to discuss using *truth tables*, and important tool in logic. We turn to that topic next.

[6] Note that this question is restricted to the sample, the subjects that were in the study. If we want to make claims about the larger population (people not in the sample), we'd have to deploy some inferential statistical tools that we have not covered.

Exercises

1. Suppose your friend has a coin with heads on one side and tails on the other. You do not know, however, whether the coin is a fair coin. All you are told is what the supposed odds are for heads and tails. Recall that if odds in favor are n:m, then odds *against* are just the inverse, m:n. In addition, recall that rewards are calculated by using odds against. Your friend presents you with the following table with partial information (the odds are odds against):

Face	Odds	Implied.Prob	Bet	Payout
Heads	1:3	X	$12	Z
Tails	4:1	Y	$5	W

  a. What is X?
  a. What is Y?
  a. What is reward Z?
  a. What is reward W?
  a. Is it more likely that the coin will flip heads or tails?
  a. What is the total of the implied probabilities (X+Y)?

Face	Odds	Implied.Prob	Bet	Payout
a. Is y	our frie	nd offering you	a fair	bet? Why or why not?

2. Suppose this time we have a coin with a substantial flat edge so that sometimes it spins and stays on the edge without falling over. The coin is a fair coin in that heads and tails are equally likely, but now there is also a positive probability that the coin could land and stay on its edge. Suppose that the probabilities are given by the table below. Using this information, what are the odds supposed to be if you want to make sure that the wagers are fair? And consequently, what should the rewards be as part of the payouts?

Face	Odds	Implied.Prob	Bet	Payout
Heads	X	0.4	$12	Z
Tails	Y	0.4	$5	W
Edge	V	0.2	$10	U
a. What	is X?			
a. What	is Y?			
a. What	is V?			
a. What	is rewa	rd Z?		
a. What	is rewa	rd W?		
a. What	is rewa	rd U?		

9
Probabilities and Logic

In this chapter we're going to approach probability more abstractly and make connections to logic. We're going to learn what it means when we say that a probability function is a *normalized measure over a possibility space*. There are three parts here: i) what is it for a function to be a measure, ii) what is a normalized measure, and iii) what is a possibility space. We will then use principles of logic to motivate several rules for going about calculating probabilities.

9.1 Measures

A function is a mapping from an input to an output. There can be many inputs that are mapped to one output, but to be a function the mapping cannot assign more than one output to an input. We typically think of functions as having numbers as inputs and outputs, but that doesn't have to be the case. In fact, as we'll see, probability functions will have a number as an output, but have non-numbers as inputs.

Consider Scotland. Scotland is known for its whisky. There are six regions (depending on who you ask) where whisky is distilled.

We can represent functions with tables. Let's suppose we have a function that takes distilling regions of Scotland as input and the number of distilleries as output.

Input	Output
Highlands	27
Speyside	62
Lowlands	3
Campbeltown	3
Islay	8
Islands	7

Notice that there are two inputs that get mapped onto the same output (Lowlands and Campbeltown both have 3 as their output), but no input has any more or less than one output (e.g., it's not the case that Islay has both 8 and 7 distilleries).

WHAT IT MEANS FOR A FUNCTION TO BE A MEASURE. **First**, the function has to have non-negative numbers as output. Our example meets this condition. **Second**, the input is some space that can have 'regions' or subsets. For example, we can organize our distilling regions into those that are on the mainland (the first four in our table) and those not on the mainland (i.e., Islay and Islands). Alternatively, we could group the list into those whose name starts with one of the first ten letters of the alphabet, those that start with 'S', and then the rest. The specifics here don't matter, just that different subsets are possible.

Third, and this important property is called *additivity*, is that the measure of a subset is the sum of the measures of the members that make up the subset. For example, suppose we're considering the mainland distillery regions (the first four on our list). Then the number of distilleries on the mainland is just the sum of the distilleries in the regions of Highlands, Speyside, Lowlands, and Campbeltown (95).

It doesn't really matter how we decide to group things. We could decide that the 'groups' are just the members themselves. In that case, there is just one number to consider. Or we could consider the entire input as the subset (here 'subset' doesn't indicate a strictly smaller set - in set theoretic speak, we say that a set is a subset of itself). For example, the number of distilleries in Scotland is, according to our table, 110 (i.e., the sum of each region).

Not all functions are measures. That is, *some functions fail to be additive*. For example, let's suppose we have a function from distilling regions to the proportion of distilleries owned by multinational companies (these numbers are not necessarily accurate).

Input	Output (proportion of owned by corps.)
Highlands	0.6
Speyside	0.4
Lowlands	0.3
Campbeltown	0
Islay	0.25
Islands	0.4

Now let's say we want to ask what proportion of the mainland distilleries are owned by multinational companies. We can't do this just

by looking at the table above. Adding up proportions won't work: it would lead to the absurd result that the proportion of mainland distilleries owned by international companies is 1.3, that is, 130%. But it's impossible to own more than 100% of the mainland distilleries, not to mention own more than 100% of anything! In general, proportions do not meet the Additivity property and thus cannot serve as the basis for measures. Typical examples of measures include length, area, and volume.

9.2   Normalized Measures

The 'universe' of a function is the entire collection of inputs. In our whisky example above, the universe of our function is the set of distillery regions in Scotland. We could have defined the universe of our function differently. For example, we might have a different function whose universe is the set of countries that have distilleries, including Scotland, Ireland, the United States, Canada, Japan, etc. Also, the universe of a function doesn't have to correspond to physical space. We could have a function with a universe of taste characteristics, like sweet, smoky, earthy, rich, peppery, floral, etc. These are not located in space like distilleries, but are rather properties of how a whisky is perceived by taste.

A *normalized* measure function is a measure function that gives a value of 1 to its universe. By doing so, the measure of every subset can be understood as a proportion of the universe of the function. Any measure can be 'normalized' by dividing the value of each output by the value of the whole universe. For example, we can normalize the function represented in the table below by dividing each output number by the sum of all outputs (110). That would give us the normalized values in the following table (rounded to four decimal places).

Input	Output	Normalized
Highlands	27	0.2455
Speyside	62	0.5636
Lowlands	3	0.0273
Campbeltown	3	0.0273
Islay	8	0.0727
Islands	7	0.0636

Notice that this normalized function satisfies the property of being additive. We can use the table to answer the question of what proportion of all distilleries in Scotland are on the mainland by adding up

the normalized values (keeping in mind that we will need to correct for errors from having rounded the values). This works because the sum of all the normalized values add up to 1. This is unlike the function represented in the previous table where the proportions do not sum to 1.

Normalizing is an important step in taking data from the world and turning it into probability functions. Sometimes the universe of a function is not very well defined. In that case, it will not be possible to normalize. So one important lesson for interpreting results of some data is to understand what the relevant universe is supposed to be.

9.3   *Possibilities and Truth Tables*

Let's look at how to build a probability function. We'll start with a really simple example. Suppose we have a fair coin and you're going to flip it twice. In the first flip there are two possible outcomes: heads (H) or tails (T). In the second flip there are again two possible outcomes: H or T. For two flips of the coin then, there are a total of four possible outcomes: HH, TH, HT, TT. We can keep track of these in a table.

	Flip 1	Flip 2
Outcome 1	H	H
Outcome 2	T	H
Outcome 3	H	T
Outcome 4	T	T

If we were to flip the coin a third time, we would have 8 possible outcomes, which we can also collect in a table

	Flip 1	Flip 2	Flip 3
Outcome 1	H	H	H
Outcome 2	T	H	H
Outcome 3	H	T	H
Outcome 4	T	T	H
Outcome 5	H	H	T
Outcome 6	T	H	T
Outcome 7	H	T	T
Outcome 8	T	T	T

If we were to flip the coin four times, there would 16 possibilities. Each time we add an additional flip, the number of possibilities dou-

bles in total. That is, where n is the number of coin flips, the total number of possible outcomes is 2^n.

Although this example is simple, it turns out to be a really fruitful way of modeling lots of other examples that have the same structure. For example, philosophers love logic and thinking about the world in terms of statements or propositions that are either true or false. Suppose we have three different propositions: P, Q, and R.

- P - "Bert drinks beer on Fridays.''
- Q - "Mandy drinks wine on Fridays.''
- R - "Florian drinks Scotch whisky on Fridays.''

For simplicity, let's assume that P is true if on the majority of Fridays Bert drinks beer, and is false otherwise. Similarly for Q and R. In the real world each of these propositions has a unique truth value. Our interest here, however, is not just what is true, but the space of possibilities. In this case, each proposition has the possibility of being true (T) or false (F). Following the coin flip example, there are in total eight possibilities. We can organize the possibilities in what logicians call a truth table.

P	Q	R
F	F	F
T	F	F
F	T	F
T	T	F
F	F	T
T	F	T
F	T	T
T	T	T

These eight possibilities provide the foundation for creating a probability function. What we first need to do is make sure that we have a measure. Let's say we give the following numbers to each row in the truth table. Where the numbers come from is not important to illustrate the point. But we can imagine, at least roughly speaking, that the universe of the function is all Fridays, and the outputs represent the fraction of times Bert, Mandy, and Florian reported what they had to drink. Notice that if we sum up the output values, we get 1. So, we have a measure, and it's a normalized measure.

Probability	P	Q	R
0.0002	F	F	F

Probability	P	Q	R
0.001	T	F	F
0.0008	F	T	F
0.008	T	T	F
0.08	F	F	T
0.1	T	F	T
0.01	F	T	T
0.8	T	T	T

We can now use this table to ask questions about the probability that a proposition is true. To do that, we look at all the rows where the proposition in question is true, and then add up the output values. For example, suppose we are interested in the probability that P is true (i.e., the probability that Bert drinks beer on a Friday). Notice that P is true in rows 2, 4, 6, and 8. So we would compute $0.001 + 0.008 + 0.1 + 0.8$ which is 0.909. That is, there is a 90.9% that Bert drinks beer on a Friday.

We aren't limited to asking the probabilities that a single proposition is true. Sometimes we'll want to ask what the probability is that both P and R is true, or that either P or R is true, or that P is false. These are all examples of more complex sentences. Logicians have a way of expressing these types of complex sentences. They call them conjunctions, disjunctions, and negations, respectively. There are some handy symbols used too:

'P∧Q' means 'P and Q' (also called conjunction) 'P∨Q' means 'P or Q' (also called disjunction) '¬P' means 'not P' (also called negation)

What's really handy about these formulations is that the truth values of the complex sentences can be fixed by the truth values of the components. For example, the rule for conjunction is

P	Q	P∧Q
T	T	T
T	F	F
F	T	F
F	F	F

As you may guess, the truth table for negation is pretty simple. The truth value is simply reversed.

P	¬P
T	F

P	¬P
F	T

For disjunction the truth value of the whole sentence is given by the following table.

P	Q	P∨Q
T	T	T
T	F	T
F	T	T
F	F	F

Notice that this is an inclusive interpretation of 'or' which means that disjunction assumes by default that when both component sentences are true then the whole sentence is true. For example, if you are asked, "Do you want ketchup or mustard on your burger" there's nothing contradictory about saying that you want both. It's easy enough to express the exclusive version 'or' by say something like 'P or Q, and not both'.[1]

We can use the truth table definitions of conjunction, disjunction, and negation to determine the possible truth values of a complex sentence. Consider for example the sentence '(P∨R) ∧ ¬Q' which can be interpreted as "On Fridays Bert drinks beer or Florian drinks Scotch whisky, but Mandy doesn't drink wine."[2] Using the rules for ∧, ∨, and ¬ we get the table below. Notice that the whole complex sentence is a conjunction, where the left conjunct is made out of a disjunction (P∨R), and the right conjunct is made out of a negation (¬Q). So we have to work from the smallest parts up to the larger parts, which means that we first evaluate'(P∨R)', then we evaluate'¬Q', and then we can do the column of truth values under '∧' (which have been boldfaced to indicate these truth values are under the main connective).

[1] Formally we would write '(P∨Q) ∧ ¬(P∧Q)'.

[2] Note that "but'' expresses"and'' with the addition of some flare to highlight a contrast.

Probability	P	Q	R	(P∨R)	∧	¬Q
0.0002	F	F	F	F	**F**	T
0.001	T	F	F	T	**T**	T
0.0008	F	T	F	F	**F**	F
0.008	T	T	F	T	**F**	F
0.08	F	F	T	T	**T**	T
0.1	T	F	T	T	**T**	T
0.01	F	T	T	T	**F**	F

Probability	P	Q	R	(P∨R) ∧ ¬Q		
0.8	T	T	T	T	**F**	F

Notice that the complex sentence is true on rows 2, 5, and 6. So as an example, if P is true but Q and R are false, then P∨R) ∧ ¬Q is true (this is row two).

If we want to calculate what the probability is that '(P∨R) ∧ ¬Q' is true, all we need to do is add up the probabilities that correspond to each row. This would be $0.001 + 0.08 + 0.1 = 0.181$. That is, there is an 18.1% chance that on Fridays Bert drinks beer or Florian drinks Scotch whisky, but Mandy doesn't drink wine.

It is worth pointing out several features about probabilities that correspond to some important logical concepts: tautologies, equivalence, entailment, and inconsistency. In fact, these will be directly connected to the axioms that are typically used to define probabilities. So it's worth paying special attention here, as we'll use these results frequently going forward.

9.3.1   Tautologies

A tautology is a sentence that is true across every possibility. The simplest example of this is a sentence 'P∨ ¬P' which says 'Either Bert drinks beers on Fridays or he doesn't'. Yes, there are complications regarding what it means to satisfy the property 'drinks beer on Fridays' - does it have to be at least more than half of all Fridays? Note that whatever line or standard you pick, that will be the same one for the negation of the sentence. So if the standard is that 'Bert drinks beers on Fridays' is true just as long as he does on 75% of all Fridays, then that sentence is false if he drinks them only on every other Friday (i.e. 50% of them).

Notice that when we take the disjunction of a sentence and the negation of that sentence, the column under the main connective (the disjunction) will be true on every row. Let's look at the truth table for 'P∨ ¬P' (notice we evaluate '¬P' first, then the disjunction):

P	P ∨ ¬P		
T	**T**	F	
F	**T**	T	

Recall that we determined the probability that 'P' is true by adding up all the probabilities in the rows were 'P' is true. If we use the same table to determine the probability that 'P' is false (i.e., we add

up all the numbers in the rows where 'P' is false) then we get the following probabilities:

Probabilities	P	P ∨ ¬P
0.909	T	**T** F
0.091	F	**T** T

Now if we add up all the rows where 'P∨ ¬P' is true, which are 1 and 2 (and there are no other rows) we get a total of 1. This makes sense upon some reflection. A tautology is always true, i.e., it is true in every possibility. We also said that a probability is a normalized measure, where the universe of the function added to 1. Since a tautology has us adding up the probabilities across all the rows, it's not surprising that they sum to 1. In brief, the probability that a tautology is true is 1.

9.3.2   *Equivalence*

Some statements are equivalent. What we mean by that is that they say the same thing, even if they differ in how they say it. For example, the sentence 'Schnee ist weiss' expresses the same proposition that is said in 'snow is white'.

In symbolic logic we can use truth tables to check whether two sentences are equivalent. Consider the sentences '¬(P ∧ Q)' and '¬P ∨ ¬ Q'. The first one reads as 'It's not the case that both Bert drinks beer on Fridays and drinks wine'. Notice that the negation is being applied to the conjunction. The second sentence reads 'Either it's not the case that Bert drinks beer on Fridays or it's not the case that Bert drinks wine on Fridays'. Notice that this second sentence is a disjunction that has two negations as component sentences.

In order to test whether two sentences are equivalent, we build a truth table and evaluate each sentence. Then we check to see whether the columns under the main connective of the sentences are identical. If they are, that means that in every possibility their truth values match, i.e., they are equivalent. When one sentence is true, so is the other, and when one is false, so is the other. When we look at the truth tables for '¬(P ∧ Q)' and '¬P ∨ ¬ Q' we see that they are indeed equivalent:

P	Q	¬(P ∧ Q)	¬P ∨ ¬ Q
T	T	F	F
T	F	T	T
F	T	T	T

P	Q	¬(P ∧ Q)	¬P ∨ ¬ Q
F	F	T	T

9.3.3   Validity and Entailment

One of the most important concepts in logic is that of **validity**. Validity is a property of arguments. An argument is a series of statements, where one is designated the conclusion and the other statements the premises (these are intended to support the conclusion). An argument is valid when *if* the premises are true, then the conclusion *has* to be true. Put differently: an argument is valid when there is no possible way of making the conclusion false and the premises true. One more way of putting it: there is no counterexample that makes the premises true but the conclusion false.

Typically arguments have two or more premises. Modus Ponens, for example, is a style of argument that has a conditional as one statement and a second statement that affirms the antecedent of the conditional. For example: If John lives in Idaho, then he lives in the US. John does indeed live in Idaho. Therefore, John lives in the US.

Not all arguments have to have two or more premises, however. Some arguments can have no premises at all! Tautologies are examples where if you make them the conclusion of an argument with no premises, the argument is still, strictly speaking, valid. (You can't make the conclusion false! So there's no counterexample.)

When an argument is valid, we say that the premises *entail* the conclusion. In many cases we'll look at arguments with just one premise and one conclusion. So if A is the premise, B the conclusion, and we have a valid argument, then we say that A entails B.

Truth tables can be used to check for entailment. What we do is make sure there is no counterexample. Put differently, we make sure that in every row where the premises are true, the conclusion is also true.

Entailment is an important concept to logicians because it preserves truth. That is, if the premises of an argument are true and you proceed through a sequence of inferences, where each inferential step is a valid one, that there is never a loss of truth. You'll never go from true premises to a false conclusion in a valid argument.

Something similar is the case if we think of probabilities instead of truth. If A entails B, then the probability of B is at least as high as the probability of A. We have this feature because when A entails B, that means that B has to be true in all the possibilities where A is (if that weren't the case, we'd have a counterexample). So if B is true in at least all the same places where A is, then the probability of B has

to be at least as great as A.

9.3.4   Inconsistency (or Mutual Exclusivity)

A standard light switch is either on or off – it's not both on and off at the same time. Now imagine that you have two light switches arranged so that when one is on it automatically turns the other one off (and vice versa). In this arrangement, the light switches are never both on at the same time, nor both off at the same time.

   Similarly, when we say that two propositions are inconsistent, we mean that whenever one is true the other is false, and when one is false the other is true. Consider '¬(P ∧ Q)' and 'P ∧ Q'. When we complete the truth tables for these, we see that they have opposite truth values on each row.

P	Q	¬(P ∧ Q)	P ∧ Q
T	T	F	T
T	F	T	F
F	T	T	F
F	F	T	F

   Whenever two propositions are inconsistent, then the probability of the disjunction of those two propositions is the sum of each of them. For example, suppose that A and B are inconsistent (that is, suppose they are *mutually exclusive*). Then $Pr(A \lor B) = Pr(A) + Pr(B)$.

9.3.5   Logic Probabilities

So let's think again about why the probability of a tautology is 1. This is actually because of two features. First, A and ¬A are inconsistent, so the probability of their disjunction (i.e. $Pr(A \lor \neg A)$) is going to be the sum of the probabilty of A and the probablity of ¬A. Second, whatever the probability of A is, the probability of '¬A' is going to be $1 - Pr(A)$. This is because inconsistent propositions cannot have probabilities that vary independently. If the probability of A is fixed, this automatically fixes the probability of '¬A' (and vice versa). And since '¬A' will cover all the remaining possibilities that 'A' did not cover and the probability of the universe (i.e., all possibilities) is 1, the probability of '¬A' is $1 - Pr(A)$.

   What about the case where A and B are not inconsistent? Is the sum of their probabilities somehow connected to the features we have been considering so far? Yes.

   Let's look at the following truth table. Notice that A and B are not inconsistent, since there are rows where their truth values match. In

the left most column we have variables representing the probability of each row. If A and B meant 'lands heads' and 'lands tails' respectively, this exercise would be much easier (since each row has a probability of 0.25). But we're asking about whether there's some general pattern that we can express even if we don't know the probabilities of any row.

Probability	A	B	A∧B	A∨B
x_1	T	T	T	T
x_2	T	F	F	T
x_3	F	T	F	T
x_4	F	F	F	F

We already know how to calculate the following probabilities:

1. $Pr(A) = x_1 + x_2$ since these are the rows where A is true.
2. $Pr(B) = x_1 + x_3$ since these are the rows where B is true.
3. $Pr(A{\wedge}B) = x_1$ since A∧B is true on just the first line
4. $Pr(A{\vee}B) = x_1 + x_2 + x_3$ since A∨B is true on lines 1-3.

The sum of the probabilities of A and B is then

$$Pr(A) + Pr(B) = x_1 + x_2 + x_1 + x_3$$

If we add lines 3 and 4, we get

$$Pr(A \wedge B) + Pr(A \vee B) = x_1 + x_1 + x_2 + x_3$$

Notice that the sum of the probabilities is actually the same (after rearranging the order): $x_1 + x_1 + x_2 + x_3$. And so we have the result that

$$Pr(A) + Pr(B) = Pr(A \wedge B) + Pr(A \vee B)$$

This is an important result, one that we will use often. Here's a different way of thinking about it, using an illustration. Let's ask how to calculate $Pr(A{\vee}B)$. As a first step, we would just add the probabilities of A and B together, i.e., $Pr(A) + Pr(B)$. However, we don't know that A and B are inconsistent, so it's possible that A and B could be true at the same time. So we need to make sure not to double count the intersection. That is, the probability that both A and B are true is already accounted for in $Pr(A{\vee}B)$, so when we add the probabilities of A and B together, we need to subtract out the probability that both are true at the same time. That what's going on when we rearrange the result we obtained above to get the following.

$$Pr(A \vee B) = Pr(A) + Pr(B) - Pr(A \wedge B)$$

9.4   *Independence*

There's one additional concept that we'll make use of going forward, but it isn't really a concept of deductive logic. Recall that two propositions are exclusive (inconsistent) when they cannot be true at the same time. So, if we flip a fair coin, landing Heads on the first toss, H_1 is exclusive of landing Tails on the first toss, T_1. But landing Heads on the first toss, H_1 is not inconsistent with landing Tails *on the second* toss, T_2. In fact, if the coin is fair and nothing about the first toss influences the outcome of the second toss, then we say that the two tosses are *independent*.

INDEPENDENCE is an important concept of inductive logic. It says that the truth of one proposition *does not change the probability* of another proposition being true. It's important not to confuse independence with exclusivity.[3] We'll have to wait until we have the concept of conditional probability to formally define independence, but a rough intuition is good enough for now.

    Independence allows us to introduce a rule for how to *multiply* probabilities. For example, suppose we have a sequence of independent tosses of our fair coin. How many possible sequences are there of two tosses? We can express them using our logic of propositions:

- $H_1 \,\&\, H_2$,
- $T_1 \,\&\, T_2$,
- $H_1 \,\&\, T_2$,
- $T_1 \,\&\, H_2$.

    How do we calculate the probability of these complex propositions? For example, how do we figure out the number $Pr(H_1 \,\&\, H_2)$ is equal to?

    Because the coin is fair, we know $Pr(H_1) = 1/2$ and $Pr(H_2) = 1/2$. The probability of heads on any given toss is always $1/2$, no matter what came before. To get the probability of $H_1 \,\&\, H_2$ we compute:

$$
\begin{aligned}
Pr(H_1 \,\&\, H_2) &= Pr(H_1) \times Pr(H_2) \\
&= 1/2 \times 1/2 \\
&= 1/4.
\end{aligned}
$$

And this is indeed correct, but *only because the coin is fair and thus the tosses are independent*. In general, we have the following as general rule of probability.

The Multiplication Rule If A and B are independent, then $Pr(A \,\&\, B) = Pr(A) \times Pr(B)$.

[3] Note that if A and B are exclusive, then finding out that A is true *does* change the probability that B is true, in fact we know it would ahve to be false! Hence, A and B are not independent.

Since our coin tosses are independent, we can multiply to calculate $Pr(H_1 \& H_2) = 1/4$. And the same reasoning applies to all four possible sequences:

$$Pr(H_1 \& H_2) = 1/4,$$
$$Pr(T_1 \& T_2) = 1/4,$$
$$Pr(H_1 \& T_2) = 1/4,$$
$$Pr(T_1 \& H_2) = 1/4.$$

IT's CRUCIAL to recognize that the multiplication rule only applies to independent propositions. You'll get the wrong answer if you're not careful about this. We'll show shortly the kinds of mistakes one might make and how to deal with propositions that aren't independent.

9.5   Summary

We've covered a lot of ground. Beware of simply trying to memorize the following key takeaways. Knowing *why* we have the following rules will be massively helpful in avoiding common errors of probabilistic reasoning (some of which we already saw with the Dutchbook argument).

- $Pr(T) = 1$ for every tautology T.
- $Pr(C) = 0$ for every contradiction C.
- $Pr(A) = Pr(B)$ if A and B are logically equivalent.
- $Pr(A \lor B) = Pr(A) + Pr(B)$, if A and B are mutually exclusive (i.e. inconsistent).
- $Pr(A \land B) = Pr(A) \times Pr(B)$, if A and B are independent.

9.6   Exercises

1. Proportions to Probabilities

    Consider again our table about the proportion of distilleries that are owen by multinational companies (these are the first four rows):

Input	Output (proportion owned by corps.)
Highlands	0.6
Speyside	0.4
Lowlands	0.3
Campbeltown	0
Islay	0.25

Input	Output (proportion owned by corps.)
Islands	0.4

We said that we can't determine what proportion of mainland distilleries are owned by multinational companies just by looking at the proportions for each region. Explain what information you would need and how you would go about figuring this out.

2. What is the probability of each of the following propositions?

   a. $A \wedge (B \wedge \neg A)$
   b. $\neg (A \wedge \neg A)$

3. Give an example of each of the following.

   a. Two statements that are mutually exclusive.
   b. Two statements that are independent.

4. For each of the following, say whether it is true or false.

   a. If propositions are independent, then they must be mutually exclusive.
   b. Independent propositions usually aren't mutually exclusive.
   c. If propositions are mutually exclusive, then they must be independent.
   d. Mutually exclusive propositions usually aren't independent.

5. Assume $Pr(A \wedge B) = 1/3$ and $Pr(A \wedge \neg B) = 1/5$. Answer each of the following:

   a. What is $Pr((A \wedge B) \vee (A \wedge \neg B))$?
   b. What is $Pr(A)$?
   c. Are $(A \wedge B)$ and $(A \wedge \neg B)$ independent?

6. Suppose A and B are independent, and A and C are mutually exclusive. Assume $Pr(A) = 1/3, Pr(B) = 1/6, Pr(C) = 1/9$, and answer each of the following:

   a. What is $Pr(A \wedge C)$?
   b. What is $Pr((A \wedge B) \vee C)$?
   c. Must $Pr(A \wedge B) = 0$?

7. Suppose a fair, six-sided die is rolled two times. What is is the probability of it landing on the same number each time?

   Hint: calculate the probability of it landing on a *different* number each time. To do this, first count the number of possible ways the two rolls could turn out. Then count how many of these are "no-repeats."

Numerous exercises have been borrowed or adapted from Weisberg's *Odds and Ends*, Chapter 5. I will admit that I lost track which exercises have come from where.

8. The Addition Rule can be extended to three propositions. If A, B, and C are all mutually exclusive with one another, then

$$Pr(A \lor B \lor C) = Pr(A) + Pr(B) + Pr(C).$$

Explain why this rule is correct. Would the same idea extend to four mutually exclusive propositions? To five?

(Hint: there's more than one way to do this. You can use an Euler diagram. Or you can derive the new rule from the original one, by thinking of $A \lor B \lor C$ as a disjunction of $A \lor B$ and C.)

9. You have a biased coin, where each toss has a 3/5 chance of landing heads. But each toss is independent of the others. Suppose you're going to flip the coin 1,000 times. The first 998 tosses all land tails. What is the probability at least one of the last two flips will be tails?

10. There are 3 empty buckets lined up. Someone takes 4 apples and places each one in a bucket. The placement of each apple is random and independent of the others. What is the probability that the first two buckets end up with no apples?

11. Suppose three cards are stacked in order: jack on top, queen in the middle, king on the bottom. If we shuffle them randomly, what is the probability the queen will still be in the middle when we're done? Assume shuffling makes every possible ordering of the cards equally likely. Hint: how many ways are there to assign each card a place in the stack? How many of these have the queen in the middle?

12. The Conjunction Fallacy

Suppose 'A' means 'Linda is a feminist' and 'B' means 'Linda is a bank teller'. Consider the truth table for A and B, with corresponding probabilities (Pr) for each assignment of truth values:

Pr	A	B
0.5	T	T
0.2	F	T
0.25	T	F
0.05	F	F

- Are A and B inconsistent? Why or why not?
- Are A and B equivalent? Why or why not?
- What is $Pr(A)$?
- What is $Pr(B)$?

- What is $Pr(A \wedge B)$?

13. Atomic vs Conjunctions

   Which of the following is true regarding the general relationship between the probability of a single proposition X and the probability of a conjunction that has X as a component?

   a. $Pr(X) \geq Pr(X \wedge Y)$
   b. $Pr(X) \leq Pr(X \wedge Y)$
   c. $Pr(X) = Pr(X \wedge Y)$
   d. None of the above. It depends on the probabilities given to X and Y.

14. Disjunction

   Consider the case of disjunction now. Which of the following is true regarding the general relationship between the probability of a single proposition X and the probability of a disjunction that has X as a component?

   a. $Pr(X) \geq Pr(X \vee Y)$
   b. $Pr(X) \leq Pr(X \vee Y)$
   c. $Pr(X) = Pr(X \vee Y)$
   d. None of the above. It depends on the probabilities given to X and Y.

15. Conjunctions and Disjunctions

   Let's compare conjunctions and disjunctions. Which of the following is true regarding the general relationship between the probability of a conjunction and the probability of a disjunction?

   a. $Pr(X \wedge Y) \geq Pr(X \vee Y)$
   b. $Pr(X \wedge Y) \leq Pr(X \vee Y)$
   c. $Pr(X \wedge Y) = Pr(X \vee Y)$
   d. None of the above. It depends on the probabilities given to X and Y.

16. Linda (again)

   Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which of the following is most probable?

   a. Linda is a bank teller
   b. Linda is a bank teller and not a feminist
   c. Linda is a bank teller and a feminist

10

Conditional Probabilities and Likelihoods

Assuming you have a driver's licence and the roads are relatively clear, the chances of crashing your car are pretty low. But if you're drunk, the chances of crashing are much higher. Probabilities change depending on the conditions.

We already have notation for symbolizing this idea. We use $P(A|B)$ to represent the probability that A is true *given* that B is true. For example, to say the probability of A given B is 30%, we write:

$$P(A|B) = 0.3$$

When we condition probabilities in this way, we call them *conditional probabilities*. Conditional probabilities play a central role in the remaining material, so let's spend some time learning how to calculate them.

10.1 Calculating Conditional Probability

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
```

Suppose I roll a fair, six-sided die behind a screen. You can't see the result, but I tell you it's an even number. What's the probability it's also a "high" number: either a 4, 5, or 6?

Maybe you figured the correct answer: 2/3. But why is that correct? Because, out of the three even numbers (2, 4, and 6), two of them are high (4 and 6). And since the die is fair, we expect it to land on a high number 2/3 of the times it lands on an even number.

This hints at a formula for $P(A|B)$.

Conditional Probability

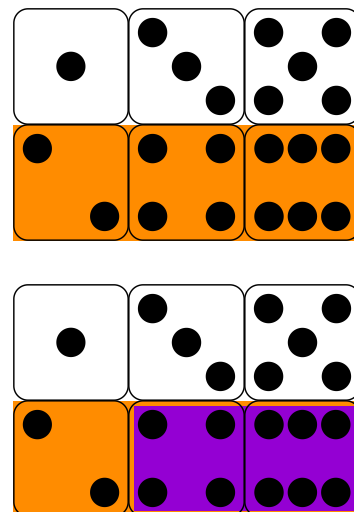$$P(A|B) = \frac{P(A \wedge B)}{P(B)}.$$

Figure 10.1: Conditional probability in a fair die roll

In the die-roll example, we considered how many of the B possibilities were also A possibilities. Which means we divided $P(A \wedge B)$ by $P(B)$.

In fact, this formula is our official definition for the concept of conditional probability. When we write the sequence of symbols $P(A|B)$, it's really just shorthand for the fraction $P(A \wedge B)/P(B)$.

```
## Warning: The 'size' argument of 'element_rect()' is deprecated as o
## i Please use the 'linewidth' argument instead.
```

In terms of an Euler diagram (Figure 10.2), the definition of conditional probability compares the size of the purple $A \wedge B$ region to the size of the whole B region, purple and blue together. If you don't mind getting a little colourful with your algebra:

$$P(A|B) = \frac{\blacksquare}{\blacksquare + \blacksquare}.$$

Figure 10.2: Conditional probability is the size of the $A \wedge B$ region compared to the entire B region.

So the definition works because, informally speaking, $P(A \wedge B)/P(B)$ is the proportion of the B outcomes that are also A outcomes.

DIVIDING by zero is a common pitfall with conditional probability. Notice how the definition of $P(A|B)$ depends on $P(B)$ being larger than zero. If $P(B) = 0$, then the formula

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

doesn't even make any sense. There is no number that results from the division on the right hand side.[1]

In such cases we say that $P(A|B)$ is *undefined*. It's not zero, or some special number. It just isn't a number.

The comedian Steven Wright once quipped that "black holes are where God divided by zero."

[1] There are alternative mathematical systems of probability, where conditional probability is defined differently to avoid this problem. But we'll stick to the standard system. In this system, there's just no such thing as "the probability of A given B" when B has zero probability.

10.2   *Application: Monty Hall Problem*

Here we'll show how the concept of conditional probability allows us to solve the Monty Hall problem in the introduction. Here's the problem description:

> On the show there are three doors (A, B, and C), one of which with a prize behind it. You get to pick one of the doors. Let's say you pick A. The host now opens one of the other two doors that you did not pick. But of course, the host doesn't want to give away the game, so the door they open will be empty. After opening one of the two doors (B or C) the host asks, do you want to switch your choice or stick with your current choice of A?

The intuitive answer, one that many mathematicians and statisticians gave at the time, is that you should be indifferent between
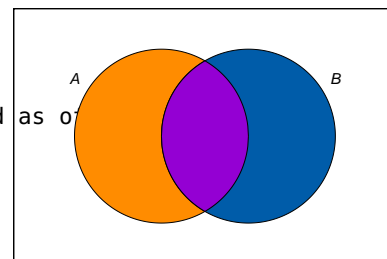
switching and staying with your choice of door A. Why? Because, the (incorrect) reasoning goes, there's two doors (A and whichever one the host didn't open) and an even chance between them of where the prize is. Notice that this question is about an *unconditional* probability.

The problem with this reasoning is that it ignores the events that proceeded. The reasoning would be apt if the game show had you picking between just two doors from the very start, and just because the host reveals what's behind the door you didn't pick, they ask you if you want to change your mind. But that's not what's going on the in the Monty Hall problem. The real question is: should you switch your choice from A *given that the host opened a non-prize door after your initial choice*? Notice that this question is about a *conditional* probability.[2]

If the prize is behind door A (the door you initially picked), then the host has a choice between opening up door B or door C. But if the prize isn't behind door A, then the host is constrained. If the prize is behind door B, then the host will open C. If the prize is behind C, then the host will open B. What we're reasoning about here are *paths* of possible events.

The first kind of event is random, it's just about the location of the prize behind one of the three doors. So the probability of the prize being behind door A is 1/3, and similarly for doors B and C. In other words, your initial guess of door A has a 1/3 chance of being right.

A diagram can help (see Stage 1). Each arrow is a branch and has a probability associated with it.

The second kind of event is the host's reveal of a non-prize door, which is *not* random *if* your choice of door A is *incorrect*, and is random if your choice of A is correct. Let's build on the diagram above, showing what the host's options are (see Stage 2). When the prize is behind door A (your guess) the host has two options, so we have two branches, each with a probability of 1/2. When the prize is behind door B, there's only one thing the host can do (so that branch has a probability of 1, which we don't bother labeling). Same thing when the prize is behind door C.

Now imagine you could play this game over and over again (thousands if you like!), always making your first pick door A.[3] One third of the time the prize will be behind door A, and of those times, half of them the host opens door B and the other half door C. So that means that the top path where the prize is behind door A and the host opens door B will happen 1/6th of the time. Similarly for the path where the prize is behind door A and the host opens C. For the other two possibilities where the prize is behind door B or door C, the host only has one option, so each of those paths will happen

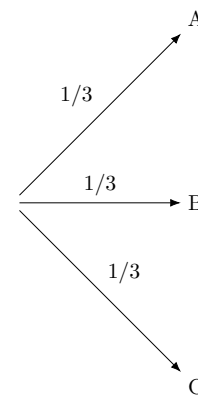[2] What the host does is *not independent* of your initial guess.
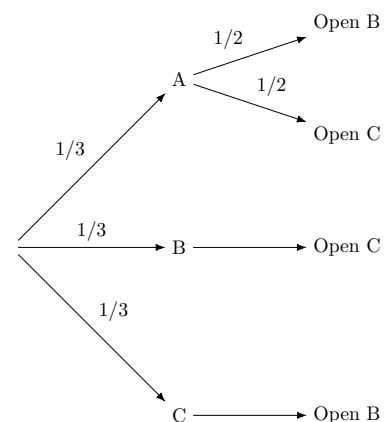

Figure 10.3: Stage 1 of tree diagram

Figure 10.4: Stage 2 of tree diagram

[3] You can try it for yourself HERE.

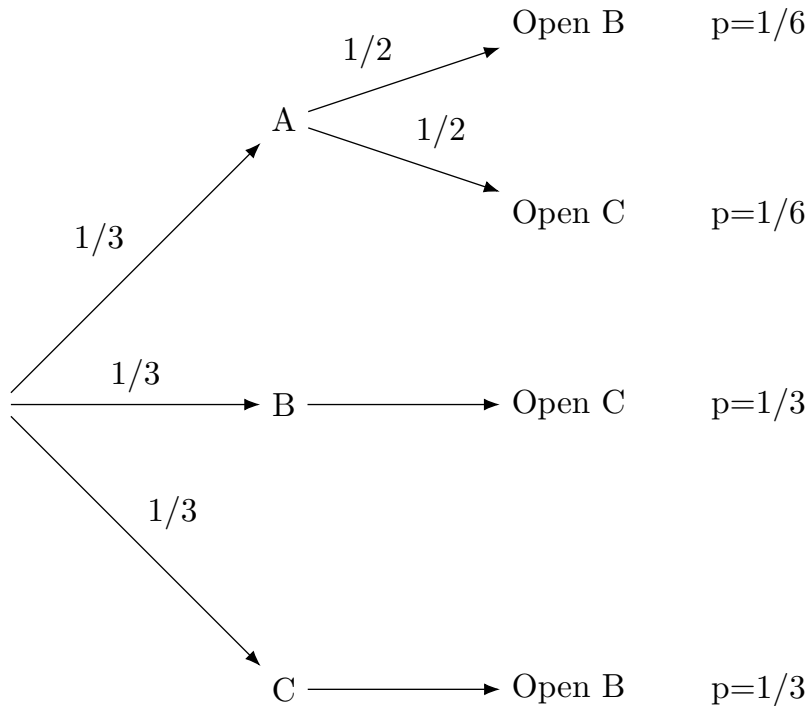1/3rd of the time. Here's our completed diagram with the probabilities at the end of the paths.

Open B   p=1/6

1/2

A

1/2

Open C   p=1/6

1/3

1/3

B ──────→ Open C   p=1/3

1/3

C ──────→ Open B   p=1/3

Of course you don't actually know which path you are on, but here's what you do know: When the host opens door C, you are more likely to be on the middle path (1/3) than you are on the path above it (1/6). On the middle path the prize is behind door B, while in the path above the prize is behind door A. Since the middle path is the more likely one you're on, you should switch your choice from A to B! The same reasoning applies if the host had opened door B instead - then you should switch your choice to C.

Notice that the probabilities at the end of the paths have to taken into account all the branches that lead to that end point. So the probability of being on the branch that goes from A to Open B is *conditional* on the previous branch that goes to A.

Here's another example.[4]

Suppose there are two urns of coloured marbles.

- Urn X contains 3 black marbles, 1 white.
- Urn Y contains 1 black marble, 3 white.

We flip a fair coin to decide which urn to draw from, heads for Urn X and tails for Urn Y. Then we draw one marble at random. The following tree diagram shows us the possible sequences of events.

[4] Thanks to Weisberg's *Odds and Ends*.

B   3/8

3/4

H

1/4
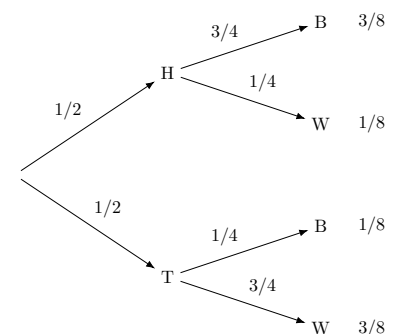
1/2

W   1/8

1/2

1/4

B   1/8

T

3/4

W   3/8

Figure 10.6: Coin Flip and Urn Draws

The probability of drawing a black marble on the top path is 3/4 because we are assuming the coin landed heads, and thus we're drawing from Urn X. If the coin lands tails instead, and we draw from Urn Y, then the chance of a black marble is instead 1/4. So these quantities are conditional probabilities:

$$Pr(B|H) = 3/4,$$
$$Pr(B|T) = 1/4.$$

Notice, though, the first branch in a tree diagram is different. In the *H*-vs.-*T* branch, the probabilities are *un*conditional, since there are no previous branches for them to be conditional on.

10.3 Independence

WE CAN STATE INDEPENDENCE in a rigorous way now. Informally stated, we said that A and B are independent when the probability of B (or the truth of B) does not affect the probability of A. The concept of conditional probability allows us to say this explicitly using probability theory:

Independence   A is independent of B if $Pr(A|B) = Pr(A)$ and $Pr(A) > 0$.

We will see more of this concept of independence, as it plays an important role in calculating probabilities. But first, there is an important kind of conditional probability called a *likelihood* that is worth taking some time to explain.

10.4 Likelihoods

ORDER MATTERS when it comes to conditional probabilities. For example, given that someone is a university student (S), the probability that they are below 40 years old (F) is relatively high. However, given that someone is below 40 years old, the probability that they are a university student is low. In symbols, $Pr(S|F) \neq Pr(F|S)$.[5]

HERE'S ANOTHER EXAMPLE.[6] Suppose a university has 10,000 students. Each is studying under one of four broad headings: Humanities, Social Sciences, STEM, or Professional. Under each of these categories, the number of students with an average grade of A, B, C, or D is listed in the following table. What is the probability a randomly selected student will have an A average, given that they are studying either Humanities or Social Sciences?

[5] There are some cases where the two are the same, but that's rare.

[6] Borrowed from Weisberg's *Odds and Ends*, chapter 6.

	Humanities	Social Sciences	STEM	Professional
A	200	600	400	900
B	500	800	1600	900
C	250	400	1500	750
D	50	200	500	450

$$Pr(A \mid (H \vee S)\,) = \frac{Pr(A \wedge (H \vee S))}{Pr(H \vee S)}$$
$$= \frac{800/10,000}{3,000/10,000}$$
$$= 4/15.$$

What about the reverse probability, that a student is studying either Humanities or Social Sciences given that they have an A average?

$$Pr((H \vee S) \mid A) = \frac{Pr((H \vee S) \wedge A)}{Pr(A)}$$
$$= \frac{800/10,000}{2,100/10,000}$$
$$= 8/21.$$

Notice how we get a different number now.

WHEN WE REASON ABOUT HYPOTHESES the order is particularly important to get right, so much so that there is a particular term that scientists will use to indicate the direction of the conditional probability. This is called a *likelihood*. To understand it, we need to make an important distinction between a hypothesis (or theory) and evidence (or data or observation).

FOR EXAMPLE, suppose a company BestShoes claims that 96% of their shoes outlast their competitor's average shoe mileage of 300 miles. Furthermore, lets say that a running magazine decides to test BestShoes' claim by sending out 2,400 shoes from BestShoes to the magazine's subscribers. The magazine finds that 133 of the shoes did not make it past 300 miles before falling apart. The *hypothesis* is the claim that each shoe has a 0.96 probability of making it past 300 miles. The *evidence* is the claim that 2,277 of 2,400 shoes made it past 300 miles.[7]

It will be convenient for us to use the letter H when talking about some hypothesis and E for some statement of evidence.

In general, a *hypothesis* is a statement about what is (or will be) the case. Our beliefs about the world are similar to scientific hypotheses (though perhaps not as technically stated) in the following sense:

[7] Later we'll see how to go about quantifying the degree of support (or lack there of) that the evidence gives to the hypothesis.

- i) there is the content of our belief, which is the statement (e.g. that BestShoes outlast their competitors), and then there is

- ii) the degree of confidence that the statement is true, which we previously learned to (indirectly) measure with the concept of fair bets. We have been representing this degree of confidence as a (personal) probability that a proposition is true, e.g. $Pr(H) = 0.75$ means a person is willing to take 3:1 odds in favor that H is true.

The idea of *confidence* as a degree of belief is not to be confused with *reasons* for that level of confidence. *Evidence* is a kind of reason that can be used to *support* a hypothesis. Evidence gives us reasons for believing that a hypothesis is true: more evidence should make us more confident.

The relationship between hypothesis and evidence can be tricky, in part because there are two different directions that are not equivalent. Recall that most of the time $P(A|B) \neq P(B|A)$. That same lesson holds for hypothesis and evidence: $P(H|E) \neq P(E|H)$.[8]

Consider the following two conditional statements:

1. If we have the body of evidence E, how probable is it that hypothesis H is true?
2. If hypothesis H is true, then how likely is it that we would see the body of evidence E?

It is a bit of curious fact that a great deal of statistics focuses on developing tools for understanding claim (2) when what we typically want to know is claim (1). For example, in most statistics classes that focus on hypothesis testing, the running intuition goes something like this:

> If the hypothesis we're testing is true, then how unusual would this result (the evidence) be?

The core idea behind all the statistical machinery (specifically what's called *frequentist* statistics) is that if the evidence would be unusual below some threshold level (called α - "alpha") *under the hypothetical assumption that the hypothesis were true*, then that supports or gives us reason to think that the hypothesis is true. If that sounds confusing, it's because it is. What we're trying to say is something like, "if the hypothesis were true, then this evidence we're seeing would be *too much of a coincidence*". If that doesn't help, you're not alone. Even scientists and statisticians have been found to misunderstand the idea.[9]

The point we need to make for our purposes is this. When scientists talk about likelihoods, they don't mean just any kind of probability. What they mean is as follows.

[8] In words: the probability that a hypothesis is true *given* that the evidence is true *is not equal to* the probability that the evidence is true given that the hypothesis is true.

[9] See https://jonathanweisberg.org/vip/significance-testing.html#significance-testing for an excellent introduction with examples of hypothesis testing.

A LIKELIHOOD is a conditional probability of seeing some evidence given the assumption that a hypothesis is true, i.e. $P(E|H)$.

10.5 Application: The Taxi Cab Problem

If there's anything to take away from this chapter about how conditional probabilities are used in (scientific) reasoning, it's to make sure you remember that $P(E|H) \neq P(H|E)$.

To ILLUSTRATE this point, consider the following famous taxi cab problem.

A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

1. 85% of the cabs in the city are Green and 15% are Blue.
2. A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

What is the probability that the cab involved in the accident was blue rather green?

Most people answer 80%, because the witness is 80% reliable. But the right answer is 12/29, or about 41%.

How could the probability be so low when the witness is 80% reliable? The short answer is: because blue cabs are rare. So most of the time, when the witness says a cab is blue, it's one of the 20% of green cabs they mistakenly identify as blue.

A diagram can help.

Imagine there are just 100 cabs in town, 85 green and 15 blue. The dashed blue line represents the cabs the witness identifies as "blue," both right or wrong. Because the witness is 80% accurate, that line encompasses 80% of the blue cabs, which is 12 cabs. But it also encompasses 20% of the green cabs, which is 17. That's a total of 29 cabs identified as "blue," only 12 of which actually are blue.

So out of the 29 cabs the witness calls "blue," only 12 really are blue. The probability a cab really is blue given the witness says so is only 12/29, about 41%.

Another way to think about the problem is that there are *two* pieces of information relevant to whether the cab is blue. The witness says the cab is blue, but also, most cabs are not blue. So there's evidence for the cab being blue, but also strong evidence against it.

The experiment was first published in 1971. It was performed by Daniel Kahneman and Amos Tversky. Their work on human reasoning reshaped the field of psychology, and eventually won a Nobel prize in 2002.
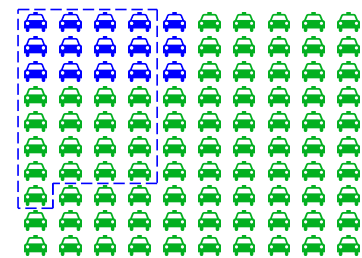

Figure 10.7: The taxicab problem. There are 15 blue cabs, 85 green. The dashed region indicates those cabs the witness identifies as "blue." It includes 80% of the blue cabs (12), and only 20% of the green ones (17). Yet it includes more green cabs than blue.

The diagram shows us how to balance these two, competing pieces of evidence and come to the correct answer.

What trips people up so much in the taxicab problem? Remember how order matters with conditional probability. In this problem, we're asked to find $Pr(B|W)$, the probability the cab is blue given that the witness says it is. That's not the same as $Pr(W|B)$, the probability the witness will say the cab is blue if it really is. The problem tells us $Pr(W|B) = 8/10$, but it doesn't tell us a number for $Pr(B|W)$. We have to figure that out.

IT'S NOT THAT THE TWO CONDITIONAL PROBABILITIES aren't related. The taxi cab problem highlights why we should keep the two distinct, but we'd be wrong to think that they aren't related at all. In fact a very famous theorem called *Bayes Theorem* shows us a fundamental connection between them. We show that next.

10.6   Exercises

1. Answer each of the following:

    Nearly all of these exercises come directly from or are versions of those in Chapter 6 of Weisberg's *Odds and Ends*.

    a. On a fair die with six sides, what is the probability of rolling a low number (1, 2, or 3) given that you roll an even number.
    b. On a fair die with eight sides, what is the probability of rolling an even number given that you roll a high number (5, 6, 7, or 8)?

2. Suppose $Pr(B) = 4/10$, $Pr(A) = 7/10$, and $Pr(B \wedge A) = 2/10$. What are each of the following probabilities?

    a. $Pr(A|B)$
    b. $Pr(B|A)$

3. Five percent of tablets made by the company Ixian have factory defects. Ten percent of the tablets made by their competitor company Guild do. A computer store buys 40% of its tablets from Ixian, and 60% from Guild.

    This exercise and the next one are based on very similar exercises from Ian Hacking's wonderful book, *An Introduction to Probability and Inductive Logic*.

    Draw a probability tree to answer the following questions.

    a. What is the probability a randomly selected tablet in the store is made by Ixian and has a factory defect?
    b. What is the probability a randomly selected tablet in the store has a factory defect?
    c. What is the probability a tablet from this store is made by Ixian, given that it has a factory defect?

4. In the city of Elizabeth, the neighbourhood of Southside has lots of chemical plants. 2% of Elizabeth's children live in Southside, and 14% of those children have been exposed to toxic levels of lead. Elsewhere in the city, only 1% of the children have toxic levels of exposure.

   Draw a probability tree to answer the following questions.

   a. What is the probability that a randomly chosen child from Elizabeth lives in Southside and has toxic levels of lead exposure?
   b. What is the probability that a randomly chosen child from Elizabeth has toxic levels of lead exposure?
   c. What is the probability that a randomly chosen child from Elizabeth who has toxic levels of lead exposure lives in Southside?

5. Imagine 100 prisoners are sentenced to death. 70 of them are housed in cell block A, the other 30 are in cell block B. Of the prisoners in cell block A, 9 are innocent. Only 1 prisoner in cell block B is innocent.

   The law requires that one prisoner be pardoned. The lucky prisoner will be selected by flipping a fair coin to choose either cell block A or B. Then a fair lottery will be used to select a random prisoner from the chosen cell block.

   What is the probability the pardoned prisoner comes from cell block A if she is innocent? Answer each of the following to find out.

   I = The pardoned prisoner is innocent.
   A = The pardoned prisoner comes from cell block A.

   a. What is $Pr(I|A)$?
   b. What is $Pr(A \land I)$?
   c. What is $Pr(I|B)$?
   d. What is $Pr(B \land I)$?
   e. What is $Pr(I)$?
   f. What is $Pr(A|I)$?
   g. Draw a probability tree to visualize and verify your calculations.

6. Suppose A, B, and C are independent, and they each have the same probability: 1/3. What is $Pr(A \land B|C)$?

7. If A and B are mutually exclusive, what is $Pr(A|B)$? Justify your answer using the definition of conditional probability.

8. Which of the following situations is impossible? Justify your answer.

   a. $Pr(A) = 1/2, Pr(A|B) = 1/2, Pr(B|A) = 1/2$.

   b. $Pr(A) = 1/2, Pr(A|B) = 1, Pr(A|\neg B) = 1$.

9. Is the following statement true or false: if A and B are mutually exclusive, then $Pr(A \lor B|C) = Pr(A|C) + Pr(B|C)$. Justify your answer.

10. Justify the second part of the "independence means multiply" mantra: if $Pr(A) > 0$, $Pr(B) > 0$, and $Pr(A \land B) = Pr(A)Pr(B)$, then A is independent of B.

    Hint: start by supposing $Pr(A) > 0$, $Pr(B) > 0$, and $Pr(A \land B) = Pr(A)Pr(B)$. Then apply some algebra and the definition of conditional probability.

> This and the remaining exercises are more challenging and ask you to think more deeply about conditional probability and independence.

11. Justify the claim that independence is symmetric: if A is independent of B, then B is independent of A.

    Hint: start by supposing that A is independent of B. Then write out $Pr(A|B)$ and apply the definition of conditional probability.

12. Suppose A, B, and C are independent. Is it possible that $Pr(A \land B \land C) = 0$? If yes, give an example where this happens. If no, prove that it cannot happen.

13. Suppose we have 4 apples and 10 buckets. We place each apple in a random bucket; the placement of each apple is independent of the others. Let B_{ij} be the proposition that apples i and j were placed in the same bucket.

    a. Is B_{12} independent of B_{34}?
    b. Is B_{12} independent of B_{23}?
    c. Is every pair of B_{ij} propositions independent?
    d. Is every trio of B_{ij} propositions independent?

14. Suppose we have a coin whose bias we want to learn, so we're going to flip it 3 times. We start out by assigning the same probability to each possible sequence of heads and tails. For example, the sequences HTH and TTT are equally likely, as are all other sequences.

    a. Before we do our 3 flips, what is the probability of HTH?
    b. What is the probability of heads on the third flip, given that the first two flips land heads?

15. Prove that if A logically entails B, then $Pr(B|A) = 1$.

16. Suppose the following three conditions hold:

   i. $Pr(A) = Pr(\neg A)$,

ii. $Pr(B|A) = Pr(B|\neg A)$,

iii. $Pr(B) > 0$.

Must the following be true then?

iv. $Pr(A|B) = Pr(A|\neg B) = 1/2$?

If yes, prove that (iv) must hold. If no, give a counterexample: draw an Euler diagram where conditions (i)–(iii) hold, but not (iv).

17. Prove that the equation $Pr(A|B)Pr(B) = Pr(B|A)Pr(A)$ always holds. (Assume both conditional probabilities are well-defined.)

18. Prove that the following equation always holds, assuming the conditional probabilities are well-defined:

$$\frac{Pr(A|B)}{Pr(B|A)} = \frac{Pr(A)}{Pr(B)}.$$

19. Does the equation $Pr(A|B) = Pr(\neg B|\neg A)$ always hold, assuming both conditional probabilities are both well-defined? If yes, prove that it does. If no, draw an eikosogram where it fails to hold.

20. Suppose an urn contains 30 black marbles and 70 white. We randomly draw 5 marbles with replacement. Let A be the proposition that 3 of the first 4 draws are black, and let B be the proposition that the 5th draw is black. Calculate the following quantity:

$$\frac{Pr(A\,|\,B)}{Pr(B\,|\,A)}\frac{Pr(B)}{Pr(A)}.$$

11

Base Rates, Priors, and Bayes Rule

The past chapter exposed you to the basic idea that underlies one of the most important theorems of probability theory. This chapter will explicitly introduce you to it: Bayes' Theorem.

11.1   *Bayes' Theorem by Example*

In the cab problem of the last chapter you were presented with two pieces of information:

1. 85% of the cabs in the city are Green and 15% are Blue.
2. A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

When asked about the probability that the cab involved in the accident was blue rather green, most of us are inclined to say 80%. The correct answer, however, is about 41%. The reason why the probability is lower than 80% is that blue cabs are rare. In other words, we failed to account for the first piece of information.[1] The reasoning we did to get to the correct answer (41%) reflects a thinking about *both* pieces of information.

[1] When we fail to incorporate information about base rates, it is known as the *base rate fallacy*.

LET'S SYSTEMATICALLY walk through this problem using our concepts of conditional and unconditional probabilities. What we want to know is the probability that the cab involved in the accident was blue (call this *H*) given that the witness said it was blue (call this *E*). That is, we want to know $Pr(H|E)$.

A BASE RATE is the information we have that is not conditioned on some other fact. In the cab problem, the base rate is contained in

(1) the number of green cabs and the number of blue ones. Another famous example of a base rate is the prevalence of some disease in a population.

A PRIOR PROBABILITY is typically used to represent some base rate. In the cab problem, for example, $Pr(H) = 0.15$ would be the prior probability of the hypothesis (H) that the cab was blue.[2]

WHAT A LIKELIHOOD does is tell us how much we should increase or decrease that prior probability given some evidence. In the taxi cab problem the evidence is the report from the witness. There are actually two steps here, but both are embedded in the information in (2) above. Let's walk through both steps in turn.

THE FIRST STEP is to remember that $Pr(E|H)$ is the probability that the witness would report the cab was blue *if the cab were in fact blue*. Given the information in (2), $Pr(E|H) = 0.8$.

THE SECOND STEP is to recognize that the witness might also report that they saw a blue cab when the cab was in fact green. This is important because the accuracy of reports (or any kind of thing that could count as evidence by "measuring" something) has two sides: how often something is reported to be the case *when in fact it is the case*, but also, how often is something reported to be the case *when in fact it is not the case*.[3] What we ultimately want here is to know how probable it is that the witness will report blue *regardless of what the color of the car is*. We can actually calculate this. The basic idea is to use two likelihoods: the probability the witness reports blue when a cab is blue ($Pr(E|H) = 0.8$), and the probability the witness reports blue when a cab is not blue, but rather green ($Pr(E|\neg H) = 0.2$). And lest we forget our previous lesson, we need to weight both of these by the frequency of blue and green cabs! In the end we then have:

$$Pr(E) = Pr(E|H) * Pr(H) + Pr(E|\neg H) * Pr(\neg H)$$

BAYES' THEOREM tells us how to combine all this information so that we can calculate $Pr(H|E)$:[4]

Bayes' Theorem

$$Pr(H|E) = \frac{Pr(H) * Pr(E|H)}{Pr(E)}$$

WHY IS THIS the way to combine the information? The are several ways we can demonstrate this. One is to use a tree diagram for the taxicab problem.[5]

[2] "Prior" because this is before we're thinking about new evidence. In the taxi cab problem, the report from the witness is the new evidence.

[3] Think of the boy who cried wolf.

[4] The theorem is named after Thomas Bayes (1701-1761) who was both a mathematician and an English minister.

[5] Let's walk through this together.

Another way to demonstrate Bayes' Theorem is to "chase" the definitions and rules we've already learned. The definition of a conditional probability is:

$$Pr(H|E) = \frac{Pr(H \wedge E)}{Pr(E)}$$

Notice that the numerator is a conjunction. To calculate the probabilities of conjunctions, we multiply the probabilites of each conjunct *if they are independent*. That is,

$$Pr(A \wedge B) = Pr(A) \times Pr(B)$$

REMEMBER THAT, if A and B are independent, then $Pr(B|A) = Pr(B)$. So if A and B are not independent, we'll need to use the conditional probability instead of the unconditional one. That is:

$$Pr(A \wedge B) = Pr(A|B) \times Pr(B)$$

In the context of hypothesis and evidence, that means we replace the numerator with:

$$Pr(H \wedge E) = Pr(E|H) \times Pr(H)$$

Plug that in to the equation above with the definition of a conditional probability, and we get:

$$Pr(H|E) = \frac{Pr(E|H) \times Pr(H)}{Pr(E)}$$

There are four distinct elements for Bayes' Theorem, and it's convenient to have names from them, most of which we've now seen:

- $Pr(H|E)$ is the *posterior* probability.
- $Pr(H)$ is the *prior* probability.
- $Pr(E|H)$ is the *likelihood*.
- $Pr(E)$ is the "normalizing constant" - it's the probability of seeing the evidence across all the possible hypotheses.

The case of $Pr(E)$ is an instance of what is known as *The Law of Total Probability*. In general, it says

Law of Total Probability If $1 > Pr(B) > 0$ then

$$Pr(A) = Pr(A|B)Pr(B) + Pr(A|\neg B)Pr(\neg B).$$

We will make use of this law enough that it's worth remembering.

11.2   *Application: Conditionalization*

THE STRENGTH of Bayes' Thereom (sometimes also known as Bayes' Rule) becomes apparent when we add a philosophical principle or rule about how we should update our beliefs as new evidence is presented to us over time. The rule, in its simplest form, is that "yesterday's posterior probability should be today's prior probability", that is:

Simple Conditionalization

$$Pr_{new}(H) = Pr_{old}(H|E)$$

Stated rigorously in words: suppose you don't know if E is true but I ask you to speculate and temporarily add E to your current stock of beliefs. The number you assign to H on the speculation that E is true should be the same as the number you assign in your unconditional credence in H when the real world has provided you with E.

LET'S WORK THROUGH AN EXAMPLE. Let's say we want to know what the probability is that Jamey has some disease, D. Jamey was recently given a test for the disease and we'll assume it came back positive, T. We want to know how confident Jamey should be that she has the disease given a positive test. Should Jamey be very confident now, or perhaps just a bit more confident? There are a lot of assumptions that should go into Jamey's learning, e.g., how reliable are the tests? How frequent is the disease in the first place?

Conditionalization tells us how to relate the assumptions to these kinds of questions. To start, when we're using conditionalization, we need to be mindful of our prior information and our posterior information. We'll use 1 for the prior time (before we got the test result) and 2 for the time we got the test result. So what we're asking is what $Pr_2(D)$ is. Simple conditionalization says

$$Pr_2(D) = Pr_1(D|T)$$

Bayes' Rule tells us what the right hand term of simple conditionalization is:

$$Pr_1(D|T) = \frac{Pr_1(D) * Pr_1(T|D)}{Pr_1(T)}$$

There are three terms now for which we need some additional information.

1. We need to know what $Pr_1(T|D)$ is, i.e., what the probability is that a test comes back positive given that the patient has the

disease (recall this is also called the likelihood). Let's say that this likelihood is 80%.[6] That gives one piece of the puzzle:

$$Pr_1(D|T) = \frac{Pr_1(D) * 0.8}{Pr_1(T)}$$

2. We need to know what $Pr_1(D)$ is, i.e., what is the baseline probability of having the disease (also known as our prior probability). One way to estimate this is to find out what percentage of the relevant demographic group has the disease. Let's say that in the demographic group to which Jamey belongs, 5 of patients have the disease. So formally, $Pr_1(D) = 0.05$ (which means $Pr_1(\neg D) = 0.95$). We have another piece of the puzzle:

$$Pr_1(D|T) = \frac{0.05 * 0.8}{Pr_1(T)}$$

3. Last, we need to know $Pr_1(T)$, i.e., what is the probability that a test comes back positive. We already know that when a patient has the disease, then the test returns a positive result 80 of the time. We represent this formally as $Pr(T|D)$. But that by itself is not enough information. We also need to know what percentage of the time we get a false positive, i.e., a test result that says positive when in fact the person does not have the disease. Let's say this is 10 of the time. Or put differently, when the patient does not have the disease, then the test returns a negative result 90 of the time. Formally, we write this as $Pr(T|\neg D)$. The Law of Total Probability says

$$Pr(T) = Pr(T|D) \times Pr(D) + Pr(T|\neg D) \times Pr(\neg D)$$

Notice that we have all these numbers already! Plugging them in we get:

$$Pr(T) = 0.8 \times 0.05 + 0.1 \times 0.95 = 0.04 + 0.095 = 0.135$$

In other words, there's a 13.5 chance that a test comes back positive.

PUTTING ALL THE PIECES TOGETHER we have:

$$\begin{aligned} Pr_2(D) &= Pr_1(D|T) \\ &= \frac{Pr_1(D) * Pr_1(T|D)}{Pr_1(T)} \\ &= \frac{0.05 * 0.8}{0.135} \\ &= 0.296 \end{aligned}$$

In words: there's about a 30 chance that Jamey has the disease given the positive test. That's probably lower than you might expect, but as

we learned before, the reason is because the base rate of the disease is itself pretty low.[7]

WHAT IF JAMEY GOT A SECOND TEST? And suppose that this second test also comes back positive. Just like before, Simple Conditionalization tells us to use prior information, but this time that information includes the results from the first test (but before the second test). Remember the adage "yesterday's posteriors are today's priors". In this context, it's "the result of updating beliefs after the first positive test is now the prior for updating on the second positive test". It doesn't have the same ring, but the idea is the same:

$$Pr_3(D) = Pr_2(D|T) = \frac{Pr_2(D) * Pr_2(T|D)}{Pr_2(T)}$$

Notice: if the tests are independent, then one of three terms will stay the same, namely $Pr_2(T|D) = Pr_1(T|D) = 0.8$. So that's an easy first piece of the puzzle:

$$Pr_3(D|T) = \frac{Pr_2(D) * Pr_2(T|D)}{Pr_2(T)} = \frac{Pr_2(D) * 0.8}{Pr_2(T)}$$

Next, Simple Conditionalizing tells us to plug in an "updated" prior, which is *not* $Pr_1(D) = 0.05$, since that reflected only the base rate. Instead, what we need to plug is in $Pr_2(D) = 0.296$ - the *posterior* we calculated after the first test. We then have:

$$Pr_3(D|T) = \frac{Pr_2(D) * Pr_2(T|D)}{Pr_2(T)} = \frac{0.296 * 0.8}{Pr_2(T)}$$

The last part for $Pr_2(T)$ is a bit more subtle. It's tempting to think that $Pr_2(T)$ will also be the same as before, but if we recall how we went about calculating it, we'll notice that it contained information about base rates in the form of priors:

$$Pr(T) = Pr(T|D) \times Pr(D) + Pr(T|\neg D) \times Pr(\neg D)$$

So if we are conducting the test on Jamey again (which we are), we have to also update the probability that we'll get a positive test given the evidence we have obtained about Jamey. That is, while the likelihoods stay the same (the tests don't change their accuracy from person to person, at least we're assuming), the priors do not. So plugging in the new priors first, and then the likelihood like we had before, we get:

$$\begin{aligned}
Pr(T) &= Pr(T|D) \times Pr_2(D) + Pr(T|\neg D) \times Pr_2(\neg D) \\
&= Pr(T|D) \times 0.296 + Pr(T|\neg D) \times 0.704 \\
&= 0.8 \times 0.296 + 0.1 \times 0.704 \\
&= 0.2368 + 0.0704 \\
&= 0.3072
\end{aligned}$$

Notice that this number is higher than earlier when we didn't have any information about Jamey. This makes intuitive sense: it *should* be more likely that we'll get a positive test from Jamey since we've accumulated some evidence that he has the disease.

So now that gives us all the puzzle pieces:

$$Pr_3(D|T) = \frac{Pr_2(D) * Pr_2(T|D)}{Pr_2(T)} = \frac{0.296 \times 0.8}{0.3072} = 0.7708$$

After two positive tests, we now have a substantially higher probability that Jamey has the rare disease. It is still far from certain. In fact, it is still *lower* than the mere accuracy of the test, $Pr(T|D) = 0.8$. But his confidence is moving in that direction.[8]

[8] Exercise: will his confidence go above $Pr(T|D) = 0.8$ for some set of evidence?

11.3   *Advanced Application*

THE PROBLEM OF UNCERTAIN EVIDENCE emerges from the recognition that our observation of evidence is itself not certain.

Simple Conditionalization with Bayes' Rule makes a simplifying assumption that some authors call into question. The conditional probability, $P(H|E)$, says what the probability of H is if we were to observe evidence E. In the conditionalization process, it is assumed that when an observation is made, that the probability of the evidence statement, $P_i(E)$, which is somewhere between zero and one, is updated and changed to $P_j(E) = 1$ (where i is a prior time and j is a later time). That's the idea from going from what we previously knew, $P_i(E)$ to the process of learning, $P_i(H|E)$, to having an updated or posterior belief, $Pr_j(H)$. And in fact, it is usually written $Pr_j(H) = P_i(H|E)$. The objection is that we can never be entirely certain of any evidence. So to say that $P(E) = 1$ is to say that we observed a logical truth or tautology!

This is called the *Problem of Uncertain Evidence*. That is, any empirical observation always has the possibility of being false - that is what it means for the world to be contingent. We should therefore not assign 1 to $P(E)$. This problem is solved by Jeffrey Conditionalization.

JEFFREY CONDITIONALIZATION assigns a degree of belief that some evidential statement is true, rather than assigning it a level of certainty by giving it 1. It is similar to Bayes' theorem, but we have to do a bit more work by accounting for the probability that E is false (that is, $\neg E$ might be true):

Jeffrey Conditionalization

$$P_j(H) = P_i(H|E) \times P_j(E) + P_i(H|\neg E) \times P_j(\neg E)$$

Notice that if we set $P_j(E) = 1$ then we get Simple Conditional-ization again (the right term of the addition will cancel out because $P_j(\neg E) = 0$.

There is some debate about how well Jeffrey Conditionalization can accommodate uncertain evidence generally. We won't pursue that here, but readers looking for more advanced discussion are encouraged to start with the Stanford Encyclopedia of Philosophy entry on Bayesian epistemology: `https://plato.stanford.edu/entries/epistemology-bayesian/supplement.html#sec-jeffrey-general`

Exercises

1. Suppose there are three colors of cabs and you are given the following information:

   In the cab problem of the last chapter you were presented with two pieces of information: i) 80% of the cabs in the city are Green, 10% are Blue, and 10% are Yellow. ii) A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the three colors 80% of the time and failed 20% of the time.

   • What is the probability that the cab in the accident is Blue given the witness report?

2. Consider our medical example with Jamey.

   • If a third test comes back positive, what should his confidence be that he has the disease?
   • If the third test instead came back negative, what should his confidence be?
   • Notice that after enough positive tests, Jamey's confidence will be *higher* than the likelihood that a single test returns a positive result *given* that the person has the disease (i.e. $Pr(T|D)$). Why is that?

3. Suppose Paul is an introvert. Should you be more confident that Paul is a librarian or in sales?

   • What's the base rate of librarians? What about people in sales?
     – A brief and rough search for the US suggests that 0.05% of people are librarians and 20% are in sales.
   • What's the probability of randomly selecting an introvert in the US?

- – "Introversion" is loosely defined, so let's say 25% of people are introverts.
- What are the likelihoods of introverts given the library position? What about given the sales position?
  - – Let's say the likelhood of introversion of a librarian is 0.8 and for a sales person it's 0.01.
- Use Bayes' Theorem to calculate the probability that Paul is a librarian.
- Use Bayes' Theorem to calculate the probability that Paul is a sales person.
- Which is more likely? And how many times more?

4. Suppose you additionally learn that Paul also has an undergraduate degree. Should you be more confident that Paul is a librarian or in sales?

- State what, if anything, you can use from the previous calculations.
- What information do you need to collect to answer this question?
- Collect some rough estimates of the information you need like we did in the above question about Paul being an introvert.
- Again, use Bayes' Theorem to calculate the probability that Paul is a librarian.
- Again, use Bayes' Theorem to calculate the probability that Paul is a sales person.
- Which is more likely given that you know Paul is both introverted and has an undergraduate degree? And how many times more?

5. Recall an earlier problem we faced:

   Five percent of tablets made by the company Ixian have factory defects. Ten percent of the tablets made by their competitor company Guild do. A computer store buys 40% of its tablets from Ixian, and 60% from Guild.

   Use Bayes' theorem to find $Pr(I|D)$, the probability a tablet from this store is made by Ixian, given that it has a factory defect?

6. Recall another problem we faced:

   In the city of Elizabeth, the neighbourhood of Southside has lots of chemical plants. 2% of Elizabeth's children live in Southside, and 14% of those children have been exposed to toxic levels of lead. Elsewhere in the city, only 1% of the children have toxic levels of exposure.

Use Bayes' theorem to find $Pr(S|L)$, the probability that a randomly chosen child from Elizabeth who has toxic levels of lead exposure lives in Southside?

7. The probability that Nasim will study for her test is 4/10. The probability that she will pass, given that she studies, is 9/10. The probability that she passes, given that she does not study, is 3/10. What is the probability that she has studied, given that she passes?

8. At the height of flu season, roughly 1 in every 100 people have the flu. But some people don't show symptoms even when they have it: only half the people who have the virus show symptoms.

Flu symptoms can also be caused by other things, like colds and allergies. So about 1 in every 20 people who don't have the flu still have flu-like symptoms.

If someone has flu-like symptoms at the height of flu season, what is the probability that they actually have the flu?

9. There is a room filled with two types of urns.

- Type A urns have 30 yellow marbles, 70 red.
- Type B urns have 20 green marbles, 80 yellow.

The two types of urn look identical, but 80% of them are Type A. You pick an urn at random and draw a marble from it at random.

a. What is the probability the marble will be yellow?

Now you look at the marble: it is yellow.

b. What is the probability the urn is a Type B urn, given that you drew a yellow marble?

10. A company makes websites, always powered by one of three server platforms: Bulldozer, Kumquat, or Penguin. Bulldozer crashes 1 out of every 10 visits, Kumquat crashes 1 in 50 visits, and Penguin only crashes 1 out of every 200 visits.

This problem is based on Exercise 6 from p. 78 of Ian Hacking's *An Introduction to Probability & Inductive Logic*.

Half of the websites are run on Bulldozer, 30% are run on Kumquat, and 20% are run on Penguin.

You visit one of their sites for the first time and it crashes. What is the probability it was run on Penguin?

11. You and Carlos are at a party, which means there's a 2/3 chance he's been drinking. You decide to experiment to find out: you throw a tennis ball to Carlos and he misses the catch. Five minutes later you try again and he misses again. Assume the two catches are independent.

When he's sober, Carlos misses a catch only two times out of ten. When he's been drinking, Carlos misses catches half the time.

What is the probability that Carlos has been drinking, given that he missed both catches?

12. The Queen Gertrude Hotel has two kinds of suites: singles have one bed, royal suites have three beds. There are 80 singles and 20 royals.

   In a single, the probability of bed bugs is 1/100. But every additional bed put in a suite doubles the chance of bed bugs.

   If a suite is inspected at random and bed bugs are found, what is the probability it's a royal?

13. Willy Wonka Chocolates Inc. makes two kinds of boxes of chocolates. The "wonk box" has four caramel chocolates and six regular chocolates. The "zonk box" has six caramel chocolates, two regular chocolates, and two mint chocolates. A third of their boxes are wonk boxes, the rest are zonk boxes. They don't mark the boxes. The only way to tell what kind of box you've bought is by trying the chocolates inside. In fact, all the chocolates look the same. You can only tell the difference by tasting them. If you buy a random box, try a chocolate at random, and find that it's caramel, what is the probability you've bought a wonk box?

14. A room contains four urns. Three of them are Type X, one is Type Y. The Type X urns each contain 3 black marbles, 2 white marbles. The Type Y urn contains 1 black marble, 4 white marbles. You are going to pick an urn at random and start drawing marbles from it at random *without* replacement. What is the probability the urn is Type X if the first draw is black?

15. Suppose I have an even mix of black and white marbles. I choose one at random without letting you see the colour, and I put it in a hat.

   Then I add a second, black marble to the hat. If I draw one marble at random from the hat and it's black, what is the probability the marble left in the hat is black?

This problem was devised by Lewis Carroll, author of *Alices Adventures in Wonderland*.

16. Suppose you have a test for some disease, which always comes up positive for people who have the disease: $Pr(P|D) = 1$. The base-rate in the population for this disease is 1%. How low does the false-positive rate $Pr(P|\neg D)$ have to be for the test to achieve 95% reliability, i.e. to have $Pr(D|P) = .95$?

17. Suppose the test for some disease is perfect for people who have the disease, $Pr(P|D) = 1$. And it's almost perfect for people who

don't have the disease: $Pr(\neg P|\neg D) = 98/99$. How high does the base rate have to be for the test to be 99% reliable, i.e. to have $Pr(D|P) = .99$?

18. An urn contains 4 marbles, either blue or green. The number of blue marbles is equally likely to be $0, 1, 2, 3$, or 4. Suppose we do 3 random draws with replacement, and the observed sequence is: blue, green, blue. What is the probability the urn contains just 1 blue marble?

12
Learning and Motivated Reasoning