Philosophy 424
LOGIC OF DECISION

Brian Weatherson, Rutgers University

2011

About this Document

These are teaching notes for a course on decision theory and game theory in Fall semester, 2011, at Rutgers University. The plan is that we'll work through roughly one chapter a week, skipping chapter 5, and using the class the week before Thanksgiving to catch up if we fall behind.

Be aware that these notes are *not* original research. If they were, I'd have to cite sources for every (non-original) idea in here, and since most of the ideas aren't original, that would be a lot! I've relied on a lot of sources in putting this together, most notably:

- Wolfram Mathworld.
- Ben Polak's OpenYale Game Theory Course.
- *Playing for Real*, by Ken Binmore, Oxford University Press.

I taught a summer seminar at Arché, St Andrews, in summer 2011 from an earlier version of the game theory chapters. I'm very grateful to the participants there for copious feedback. Hopefully this version has fewer errors and more insights!

Reproduction

Citation

Since this isn't a scholarly work, it shouldn't be cited in scholarly work.

Contents

1 Introduction 1

1.1 Decisions and Games 1
1.2 Previews 3
1.3 Example: Newcomb 4
1.4 Example: Voting 5
1.5 Dominance Reasoning 6
1.6 States and Choices 6
1.7 Maximin and Maximax 8
1.8 Ordinal and Cardinal Utilities 9
1.9 Regret 10
1.10 Likely Outcomes 11
1.11 Do What's Likely to Work 12
1.12 Probability Defined 14
1.13 Measures 14
1.14 Normalised Measures 16
1.15 Formalities 17
1.16 Possibility Space 18

2 Probability 21

2.1 Compound Sentences 21
2.2 Equivalence, Entailment, Inconsistency, and Logical Truth 24
2.3 Two Important Results 26
2.4 Axioms of Probability 28
2.5 Truth Tables and Possibilities 29
2.6 Propositions and Possibilities 32
2.7 Conditional Probability 34
2.8 Bayes Theorem 36
2.9 Conditionalisation 38
2.10 Conglomerability 40

2.11 Independence 41
2.12 Kinds of Independence 42
2.13 Gamblers' Fallacy 45

3 Utility **47**
3.1 Expected Values 47
3.2 Maximise Expected Utility Rule 48
3.3 Structural Features 51
3.4 Generalising Dominance 52
3.5 Sure Thing Principle 55
3.6 Allais Paradox 57

4 Working Out Probabilities **59**
4.1 Frequency 59
4.2 Degrees of Belief 61
4.3 Credences and Norms 64
4.4 Evidential Probability 65
4.5 Objective Chances 66
4.6 The Principal Principle and Direct Inference 67
4.7 Money and Utility 69
4.8 Insurance 70
4.9 Diversification 71
4.10 Selling Insurance 72

5 Utility **75**
5.1 Utility and Welfare 75
5.2 Experiences and Welfare 76
5.3 Objective List Theories 78
5.4 Preference Based Theories 80
5.5 Interpersonal Comparisons 82
5.6 Which Desires Count 83

6 Newcomb's Puzzle **85**
6.1 Two Principles of Decision Theory 86
6.2 Arguments for Taking Both Boxes 88
6.3 Tickle Defence 93
6.4 Causal and Evidential Decision Theory 94
6.5 Weak Newcomb Problems 99

7 Introduction to Games 101
7.1 Prisoners' Dilemma 103
7.2 Dominating Strategies 107
7.3 Iterated Dominance 108
7.4 Iteration and Knowledge 113
7.5 Strong and Weak Dominance 115
7.6 Puzzles about Weak Domination 118

8 Games and Time 123
8.1 Normal Form and Extensive Form 123
8.2 Backwards Induction 127
8.3 Value of Games 130

9 Best Responses 133
9.1 Best Responses 133
9.2 Iterated Best Responses 134
9.3 Nash Equilibrium in Simultaneous Move Games 137
9.4 Mixtures and Dominance 139
9.5 What is a Mixed Strategy? 139
9.6 Nash, Best Response and Repeated Games 143
9.7 Equilibrium and Ratification 144
9.8 Game Theory as Epistemology 150

10 Finding Equilibria 155
10.1 Finding Mixed Strategy Equilibria 155
10.2 Coordination Games 159
10.3 Mixed Strategies in Multiple Equilibrium Games 162
10.4 Value of Communication 165

11 Refining Equilibria 169
11.1 Subgame Perfect Equilibrium 170
11.2 Forward Induction 172
11.3 Perfect Bayesian Equilibrium 174
11.4 Signalling Games 176
11.5 Other Types of Constraint 185

12 Backward Induction and its Discontents 189
12.1 Problems with Backwards Induction 189
12.2 Money Burning Game 191
12.3 Iterated Prisoners' Dilemma 193

13 Group Decisions 199
13.1 Making a Decision 200
13.2 Desiderata for Preference Aggregation Mechanisms 202
13.3 Ranking Functions 204
13.4 Cyclic Preferences 206
13.5 Proofs of Arrow's Theorem 208

14 Voting Systems 211
14.1 Plurality voting 212
14.2 Runoff Voting 213
14.3 Instant Runoff Voting 215
14.4 Borda Count 216
14.5 Approval Voting 218
14.6 Range Voting 220

Chapter 1

Introduction

1.1 Decisions and Games

This course is an introduction to decision theory. We're interested in what to do when the outcomes of your actions depend on some external facts about which you are uncertain. The simplest such decision has the following structure.

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

The choices are the options you can take. The states are the ways the world can be that affect how good an outcome you'll get. And the variables, a, b, c and d are numbers measuring how good those outcomes are. For now we'll simply have higher numbers representing better outcomes, though eventually we'll want the numbers to reflect how good various outcomes are.

Let's illustrate this with a simple example. It's a Sunday afternoon, and you have the choice between watching a football game and finishing a paper due on Monday. It will be a little painful to do the paper after the football, but not impossible. It will be fun to watch football, at least if your team wins. But if they lose you'll have spent the afternoon watching them lose, and still have the paper to write. On the other hand, you'll feel bad if you skip the game and they win. So we might have the following decision table.

	Your Team Wins	Your Team Loses
Watch Football	4	1
Work on Paper	2	3

The numbers of course could be different if you have different preferences. Perhaps your desire for your team to win is stronger than your desire to avoid regretting missing the game. In that case the table might look like this.

	Your Team Wins	Your Team Loses
Watch Football	4	1
Work on Paper	3	2

Either way, what turns out to be for the best depends on what the state of the world is. These are the kinds of decisions with which we'll be interested.

Sometimes the relevant state of the world is the action of someone who is, in some loose sense, interacting with you. For instance, imagine you are playing a game of rock-paper-scissors. We can represent that game using the following table, with the rows for your choices and the columns for the other person's choices.

	Rock	Paper	Scissors
Rock	0	-1	1
Paper	1	0	-1
Scissors	-1	1	0

Not all games are competitive like this. Some games involve coordination. For instance, imagine you and a friend are trying to meet up somewhere in New York City. You want to go to a movie, and your friend wants to go to a play, but neither of you wants to go to something on their own. Sadly, your cell phone is dead, so you'll just have to go to either the movie theater or the playhouse, and hope your friend goes to the same location. We might represent the game you and your friend are playing this way.

	Movie Theater	Playhouse
Movie Theater	4, 1	0, 0
Playhouse	0, 0	1, 4

In each cell now there are two numbers, representing first how good the outcome is for you, and second how good it is for your friend. So if you both go to the movies, that's the best outcome for you, and the second-best for your friend. But if you go to different things, that's the worst result for both of you. We'll look a bit at games like this where the party's interests are neither strictly allied nor strictly competitive.

Traditionally there is a large division between **decision theory**, where the outcome depends just on your choice and the impersonal world, and **game theory**, where the outcome depends on the choices made by multiple interacting agents. We'll follow this tradition here, focussing on decision theory for the first two-thirds of the course, and then shifting our attention to game theory. But it's worth noting that this division is fairly arbitrary. Some decisions depend for their outcome on the choices of entities that are borderline agents, such as animals or very young children. And some decisions depend for their outcome on choices of agents that are only minimally interacting with you. For these reasons, among others, we should be suspicious of theories that draw a sharp line between decision theory and game theory.

One reason for drawing this distinction, however, is that sometimes we need less information to 'solve' a game than we need to solve a decision puzzle. To solve a decision puzzle, we usually need to be told how likely the different possible states of the world are. That isn't always necessary for games; sometimes we can figure out how likely they are from the fact that the relevant state of the world consists of a rational choice by a player. Much later in the course, we will look at this idea

1.2 Previews

Just thinking intuitively about decisions like whether to watch football, it seems clear that how likely the various states of the world are is highly relevant to what you should do. If you're more or less certain that your team will win, and you'll enjoy watching the win, then you should watch the game. But if you're more or less certain that your team will lose, then it's better to start working on the term paper. That intuition, that how likely the various states are affects what the right decision is, is central to modern decision theory.

The best way we have to formally regiment likelihoods is **probability theory**. So we'll spend quite a bit of time in this course looking at probability, because it is central to good decision making. In particular, we'll be looking at four things.

First, we'll spend some time going over the basics of probability theory itself. Many people, most people in fact, make simple errors when trying to reason probabilistically. This is especially true when trying to reason with so-called **conditional probabilities**. We'll look at a few common errors, and look at ways to avoid them.

Second, we'll look at some questions that come up when we try to extend probability theory to cases where there are infinitely many ways the world could be. Some issues that come up in these cases affect how we understand probability, and in any case the issues are philosophically interesting in their own right.

Third, we'll look at some arguments as to why we should use probability theory, rather than some other theory of uncertainty, in our reasoning. Outside of philosophy it is sometimes taken for granted that we should mathematically represent uncertainties as probabilities, but this is in fact quite a striking and, if true, profound result. So we'll pay some attention to arguments in favour of using probabilities. Some of these arguments will also be relevant to questions about whether we should represent the value of outcomes with numbers.

Finally, we'll look a little at where probabilities come from. The focus here will largely be negative. We'll look at reasons why some simple identifications of probabilities either with numbers of options or with frequencies are unhelpful at best.

In the middle of the course, we'll look at a few modern puzzles that have been the focus of attention in decision theory. Later today we'll go over a couple of examples that illustrate what we'll be covering in this section.

The final part of the course will be on game theory. We'll be looking at some of the famous examples of two person games. (We've already seen a version of one, the movie and play game, above.) And we'll be looking at the use of **equilibrium** concepts in analysing various kinds of games.

We'll end with a point that we mentioned above, the connection between decision theory and game theory. Some parts of the standard treatment of game theory seem not to be consistent with the best form of decision theory that we'll look at. So we'll want to see how much revision is needed to accommodate our decision theoretic results.

1.3 Example: Newcomb

In front of you are two boxes, call them A and B. You call see that in box B there is $1000, but you cannot see what is in box A. You have a choice, but not perhaps the one you were expecting. Your first option is to take just box A, whose contents you do not know. Your other option is to take both box A and box B, with the extra $1000.

There is, as you may have guessed, a catch. A demon has predicted whether you will take just one box or take two boxes. The demon is very good at predicting these things – in the past she has made many similar predictions and been right every time. If the demon predicts that you will take both boxes, then she's put nothing in box A. If the demon predicts you will take just one box, she has put $1,000,000 in box A. So the table looks like this.

	Predicts 1 box	Predicts 2 boxes
Take 1 box	\$1,000,000	\$0
Take 2 boxes	\$1,001,000	\$1,000

There are interesting arguments for each of the two options here.

The argument for taking just one box is easy. The way the story has been set up, lots of people have taken this challenge before you. Those that have taken 1 box have walked away with a million dollars. Those that have taken both have walked away with a thousand dollars. You'd prefer to be in the first group to being in the second group, so you should take just one box.

The argument for taking both boxes is also easy. Either the demon has put the million in box A or she hasn't. If she has, you're better off taking both boxes. That way you'll get \$1,001,000 rather than \$1,000,000. If she has not, you're better off taking both boxes. That way you'll get \$1,000 rather than \$0. Either way, you're better off taking both boxes, so you should do that.

Both arguments seem quite strong. The problem is that they lead to incompatible conclusions. So which is correct?

1.4 Example: Voting

There are nine people on a committee, and they have to decide between three candidates, A, B and C for a job. They each have a ranking of the three candidates.

- Four members of the committee think A is best, then B, then C.
- Three members of the committee think C is best, then B then A.
- Two members of the committee think B is best, then C, then A.

Who should get the job? We can make a case for each of the three.

- A has the most votes. If the committee uses the voting method most common in American elections, i.e., to give the job to the person with the most votes, A wins.
- But no candidate has a majority. Perhaps we should, as many voting systems do, eliminate the last place candidate, and have a 'run-off' election between the top two. (That's what happens in a number of elections in the South, in most elections in Australia, and for choosing the venue for events like the World Cup or Olympics.) Since B is last, they get eliminated. Then the two people who supported B will change their vote to C, and C will win 5-4, and get the job.
- On the other hand, there is a case for B. A majority of the committee prefers B to A. And a majority of the committee prefers B to C. And

> many theorists have argued that when a candidate is preferred to each of their rivals by a majority of voters, that should be sufficient for winning.

At the end of the course, we will look at voting systems in more detail, and see which systems generally look most plausible. That might help resolve puzzle cases like this one.

1.5 Dominance Reasoning

The simplest rule we can use for decision making is *never choose dominated options*. There is a stronger and a weaker version of this rule.

An option A **strongly dominated** another option B if in every state, A leads to better outcomes than B. A **weakly dominates** B if in every state, A leads to at least as good an outcome as B, and in some states it leads to better outcomes.

We can use each of these as decision principles. The dominance principle we'll be primarily interested in says that if A strongly dominates B, then A should be preferred to B. We get a slightly *stronger* principle if we use *weak* dominance. That is, we get a slightly stronger principle if we say that whenever A weakly dominates B, A should be chosen over B. It's a stronger principle because it applies in more cases — that is, whenever A strongly dominates B, it also weakly dominates B.

Dominance principles seem very intuitive when applied to everyday decision cases. Consider, for example, a revised version of our case about choosing whether to watch football or work on a term paper. Imagine that you'll do very badly on the term paper if you leave it to the last minute. And imagine that the term paper is vitally important for something that matters to your future. Then we might set up the decision table as follows.

	Your team wins	Your team loses
Watch football	2	1
Work on paper	4	3

If your team wins, you are better off working on the paper, since 4 > 2. And if your team loses, you are better off working on the paper, since 3 > 1. So either way you are better off working on the paper. So you should work on the paper.

1.6 States and Choices

Here is an example from Jim Joyce that suggests that dominance might not be as straightforward a rule as we suggested above.

> Suppose you have just parked in a seedy neighborhood when a man approaches and offers to "protect" your car from harm for $10. You recognize this as extortion and have heard that people who refuse "protection" invariably return to find their windshields smashed. Those who pay find their cars intact. You cannot park anywhere else because you are late for an important meeting. It costs $400 to replace a windshield. Should you buy "protection"? Dominance says that you should not. Since you would rather have the extra $10 both in the even that your windshield is smashed and in the event that it is not, Dominance tells you not to pay. (from Joyce, *The Foundations of Causal Decision Theory*, pp 115-6.)

We can put this in a table to make the dominance argument that Joyce suggests clearer.

	Broken Windshield	Unbroken Windshield
Pay extortion	-$410	-$10
Don't pay	-$400	0

In each column, the number in the 'Don't pay' row is higher than the number in the 'Pay extortion' row. So it looks just like the case above where we said dominance gives a clear answer about what to do. But the conclusion is crazy. Here is how Joyce explains what goes wrong in the dominance argument.

> Of course, this is absurd. Your choice has a direct influence on the state of the world; refusing to pay makes it likly that your windshield will be smashed while paying makes this unlikely. The extortionist is a despicable person, but he has you over a barrel and investing a mere $10 now saves $400 down the line. You should pay now (and alert the police later).

This seems like a general principle we should endorse. We should define *states* as being, intuitively, independent of choices. The idea behind the tables we've been using is that the outcome should depend on two factors - what you do and what the world does. If the 'states' are dependent on what choice you make, then we won't have successfully 'factorised' the dependence of outcomes into these two components.

We've used a very intuitive notion of 'independence' here, and we'll have a lot more to say about that in later sections. It turns out that there are a lot of ways to think about independence, and they yield different recommendations about what to do. For now, we'll try to use 'states' that are clearly independent of the choices we make.

1.7 Maximin and Maximax

Dominance is a (relatively) uncontroversial rule, but it doesn't cover a lot of cases. We'll start now lookintg at rules that are more or less comprehensive. To start off, let's consider rules that we might consider rules for optimists and pessimists respectively.

The **Maximax** rule says that you should **maxi**mise the **max**imum outcome you can get. Basically, consider the best possible outcome, consider what you'd have to do to bring that about, and do it. In general, this isn't a very plausible rule. It recommends taking any kind of gamble that you are offered. If you took this rule to Wall St, it would recommend buying the riskiest derivatives you could find, because they might turn out to have the best results. Perhaps needless to say, I don't recommend that strategy.

The **Maximin** rule says that you should **maxi**mise the **min**imum outcome you can get. So for every choice, you look at the worst-case scenario for that choice. You then pick the option that has the least bad worst case scenario. Consider the following list of preferences from our watch football/work on paper example.

	Your team wins	Your team loses
Watch football	4	1
Work on paper	3	2

So you'd prefer your team to win, and you'd prefer to watch if they win, and work if they lose. So the worst case scenario if you watch the game is that they lose - the worst case scenario of all in the game. But the worst case scenario if you don't watch is also that they lose. Still that wasn't as bad as watching the game and seeing them lose. So you should work on the paper.

We can change the example a little without changing the recommendation.

	Your team wins	Your team loses
Watch football	4	1
Work on paper	2	3

In this example, your regret at missing the game overrides your desire for your team to win. So if you don't watch, you'd prefer that they lose. Still the worst case scenario is you don't watch is 2, and the worst case scenario if you do watch is 1. So, according to maximin, you should not watch.

Note in this case that the worst case scenario is a different state for different choices. Maximin doesn't require that you pick some 'absolute' worst-case scenario and decide on the assumption it is going to happen. Rather, you look at different worst case scenarios for different choices, and compare them.

1.8 Ordinal and Cardinal Utilities

All of the rules we've looked at so far depend only on the *ranking* of various options. They don't depend on how much we prefer one option over another. They just depend on which order we rank goods is.

To use the technical language, so far we've just looked at rules that just rely on **ordinal utilities**. The term *ordinal* here means that we only look at the **order** of the options. The rules that we'll look at rely on **cardinal utilities**. Whenever we're associating outcomes with numbers in a way that the magnitudes of the differences between the numbers matters, we're using cardinal utilities.

It is rather intuitive that something more than the ordering of outcomes should matter to what decisions we make. Imagine that two agents, Chris and Robin, each have to make a decision between two airlines to fly them from New York to San Francisco. One airline is more expensive, the other is more reliable. To oversimplify things, let's say the unreliable airline runs well in good weather, but in bad weather, things go wrong. And Chris and Robin have no way of finding out what the weather along the way will be. They would prefer to save money, but they'd certainly not prefer for things to go badly wrong. So they face the following decision table.

	Good weather	Bad weather
Fly cheap airline	4	1
Fly good airline	3	2

If we're just looking at the ordering of outcomes, that is the decision problem facing both Chris and Robin.

But now let's fill in some more details about the cheap airlines they could fly. The cheap airline that Chris might fly has a problem with luggage. If the weather is bad, their passengers' luggage will be a day late getting to San Francisco. The cheap airline that Robin might fly has a problem with staying in the air. If the weather is bad, their plane will crash.

Those seem like very different decision problems. It might be worth risking one's luggage being a day late in order to get a cheap plane ticket. It's not worth risking, seriously risking, a plane crash. (Of course, we all take some risk of being in a plane crash, unless we only ever fly the most reliable airline that we possibly could.) That's to say, Chris and Robin are facing very different decision problems, even though the ranking of the four possible outcomes is the same in each of their cases. So it seems like some decision rules should be sensitive to magnitudes of differences between options. The first kind of rule we'll look at uses the notion of regret.

1.9 Regret

Whenever you are faced with a decision problem without a dominating option, there is a chance that you'll end up taking an option that turns out to be suboptimal. If that happens there is a chance that you'll regret the choice you take. That isn't always the case. Sometimes you decide that you're happy with the choice you made after all. Sometimes you're in no position to regret what you chose because the combination of your choice and the world leaves you dead.

Despite these complications, we'll define the **regret** of a choice to be the difference between the value of the best choice given that state, and the value of the choice in question. So imagine that you have a choice between going to the movies, going on a picnic or going to a baseball game. And the world might produce a sunny day, a light rain day, or a thunderstorm. We might imagine that your values for the nine possible choice-world combinations are as follows.

	Sunny	Light rain	Thunderstorm
Picnic	20	5	0
Baseball	15	2	6
Movies	8	10	9

Then the amount of regret associated with each choice, in each state, is as follows

	Sunny	Light rain	Thunderstorm
Picnic	0	5	9
Baseball	5	8	3
Movies	12	0	0

Look at the middle cell in the table, the 8 in the baseball row and light rain column. The reason that's a 8 is that in that possibility, you get utility 2. But you could have got utility 10 from going to the movies. So the regret level is 10 - 2, that is, 8.

There are a few rules that we can describe using the notion of regret. The most commonly discussed one is called **Minimax regret**. The idea behind this rule is that you look at what the maximum possible regret is for each option. So in the above example, the picnic could end up with a regret of 9, the baseball with a regret of 8, and the movies with a regret of 12. Then you pick the option with the *lowest* maximum possible regret. In this case, that's the baseball.

The minimax regret rule leads to plausible outcomes in a lot of cases. But it has one odd structural property. In this case it recommends choosing the baseball over the movies and picnic. Indeed, it thinks going to the movies is the worst option of all. But now imagine that the picnic is ruled out as an option. (Perhaps we find out that we don't have any way to get picnic food.) Then we have the following table.

	Sunny	Light rain	Thunderstorm
Baseball	15	2	6
Movies	8	10	9

And now the amount of regret associated with each option is as follows.

	Sunny	Light rain	Thunderstorm
Baseball	0	8	3
Movies	7	0	0

Now the maximum regret associated with going to the baseball is 8. And the maximum regret associated with going to the movies is 7. So minimax regret recommends going to the movies.

Something very odd just happened. We had settled on a decision: going to the baseball. Then an option that we'd decided against, a seemingly irrelevant option, was ruled out. And because of that we made a new decision: going to the movies. It seems that this is an odd result. It violates what decision theorists call the **Irrelevance of Independence Alternatives**. Formally, this principle says that if option C is chosen from some set S of options, then C should be chosen from any set of options that (a) includes C and (b) only includes choices in S. The minimax regret rule violates this principle, and that seems like an unattractive feature of the rule.

1.10 Likely Outcomes

Earlier we considered the a decision problem, basically deciding what to do with a Sunday afternoon, that had the following table.

	Sunny	Light rain	Thunderstorm
Picnic	20	5	0
Baseball	15	2	6
Movies	8	10	9

We looked at how a few different decision rules would treat this decision. The maximin rule would recommend going to the movies, the maximax rule going to the picnic, and the minimax regret rule going to the baseball.

But if we were faced with that kind of decision in real life, we wouldn't sit down to start thinking about which of those three rules were correct, and using the answer to that philosophical question to determine what to do. Rather, we'd consult a weather forecast. If it looked like it was going to be sunny, we'd go on a picnic. If it looked like it was going to rain, we'd go to the movie. What's relevant is how likely each of the three states of the world are. That's something none of our decision rules to date have considered, and it seems like a large omission.

In general, how likely various states are plays a major role in deciding what to do. Consider the following broad kind of decision problem. There is a particular disease that, if you catch it and don't have any drugs to treat it with, is likely fatal. Buying the drugs in question will cost $500. Do you buy the drugs?

Well, that probably depends on how likely it is that you'll catch the disease in the first place. The case isn't entirely hypothetical. You or I could, at this moment, be stockpiling drugs that treat anthrax poisoning, or avian flu. I'm not buying drugs to defend against either thing. If it looked more likely that there would be more terrorist attacks using anthrax, or an avian flu epidemic, then it would be sensible to spend $500, and perhaps a lot more, defending against them. As it stands, that doesn't seem particularly sensible. (I have no idea exactly how much buying the relevant drugs would cost; the $500 figure was somewhat made up. I suspect it would be a rolling cost because the drugs would go 'stale'.)

We'll end this chapter by looking at a decision rule that might be employed taking account of the likelihood of various outcomes. This will start us down the track to discussions of probability, a subject that we'll be interested in for most of the rest of the course.

1.11 Do What's Likely to Work

The following decision rule doesn't have a catchy name, but I'll call it Do What's Likely to Work. The idea is that we should look at the various states that could come about, and decide which of them is most likely to actually happen. This is more or less what we would do in the decision above about what to do with a Sunday afternoon. The rule says then we should make the choice that will result in the best outcome in that most likely of states.

The rule has two nice advantages. First, it doesn't require a very sophisticated theory of likelihoods. It just requires us to be able to rank the various states in terms of how likely they are. Using some language from the previous section, we rely on a *ordinarl* rather than a *cardinal* theory of likelihoods. Second, it matches up well enough with a lot of our everyday decisions. In real life cases like the above example, we really do decide what state is likely to be actual (i.e. decide what the weather is likely to be) then decide what would be best to do in that circumstance.

But the rule also leads to implausible recommendations in other real life cases. Indeed, in some cases it is so implausible that it seems that it must at some level be deeply mistaken. Here is a simple example of such a case.

You have been exposed to a deadly virus. About 1/3 of people who are exposed to the virus are infected by it, and all those infected by it die unless they receive a vaccine. By the time any symptoms of the virus show up, it is too late for the vaccine to work. You are offered a vaccine for $500. Do you take it or not?

Well, the most likely state of the world is that you don't have the virus. After all, only 1/3 of people who are exposed catch the virus. The other 2/3 do not, and the odds are that you are in that group. And if you don't have the virus, it isn't worth paying $500 for a vaccine against a virus you haven't caught. So by "Do What's Likely to Work", you should decline the vaccine.

But that's crazy! It seems as clear as anything that you should pay for the vaccine. You're in serious danger of dying here, and getting rid of that risk for $500 seems like a good deal. So "Do What's Likely to Work" gives you the wrong result. There's a reason for this. You stand to lose a lot if you die. And while $500 is a lot of money, it's a lot less of a loss than dying. Whenever the downside is very different depending on which choice you make, sometimes you should avoid the bigger loss, rather than doing the thing that is most likely to lead to the right result.

Indeed, sometimes the sensible decision is one that leads to the best outcome in no possible states at all. Consider the following situation. You've caught a nasty virus, which will be fatal unless treated. Happily, there is a treatment for the virus, and it only costs $100. Unhappily, there are two strands of the virus, call them A and B. And each strand requires a different treatment. If you have the A strand, and only get the treatment for the B virus, you'll die. Happily, you can have each of the two treatments; they don't interact with each other in nasty ways. So here are your options.

	Have strand A	Have strand B
Get treatment A only	Pay \$100 + live	Pay \$100 + die
Get treatment B only	Pay \$100 + die	Pay \$100 + live
Get both treatments	Pay \$200 + live	Pay \$200 + live

Now the sensible thing to do is to get both treatments. But if you have strand A, the best thing to do is to get treatment A only. And if you have strand B, the best thing to do is to get treatment B only. There is no state whatsoever in which getting both treatments leads to the best outcome. Note that "Do What's Likely to Work" only ever recommends options that are the best in some state or other. So it's a real problem that sometimes the thing to do does not produce the best outcome in *any* situation.

1.12 Probability Defined

We talk informally about probabilities all the time. We might say that it is more probable than not that such-and-such team will make the playoffs. Or we might say that it's very probable that a particular defendant will be convicted at his trial. Or that it isn't very probable that the next card will be the one we need to complete this royal flush.

We also talk formally about probability in mathematical contexts. Formally, a probability function is a normalised measure over a possibility space. Below we'll be saying a fair bit about what each of those terms mean. We'll start with *measure*, then say what a *normalised measure* is, and finally (over the next two days) say something about *possibility spaces*.

There is a very important philosophical question about the connection between our informal talk and our formal talk. In particular, it is a very deep question whether this particular kind of formal model is the right model to represent our informal, intuitive concept. The vast majority of philosophers, statisticians, economists and others who work on these topics think it is, though as always there are dissenters. In this course, however, we will mostly stick to the orthodox view; it is complicated enough without worrying about variants! So we need to understand what the mathematicians are talking about when they talk about probabilities. And that requires starting with the notion of a measure.

1.13 Measures

A measure is a function from 'regions' of some space to non-negative numbers with the following property. If A is a region that divides exactly into regions B and C, then the measure of A is the sum of the measures of B and C. And more generally, if A divides exactly into regions B_1, B_2, ..., B_n, then the measure of A will be the sum of the measures of B_1, B_2, ... and B_n.

Here's a simple example of a measure: the function that takes as input any part of New York City, and returns as output the population of that part. Assume that the following numbers are the populations of New York's five boroughs. (These numbers are far from accurate.)

Borough	**Population**
Brooklyn	2,500,000
Queens	2,000,000
Manhattan	1,500,000
The Bronx	1,000,000
Staten Island	500,000

We can already think of this as a function, with the left hand column giving the inputs, and the right hand column the values. Now if this function is a *measure*, it should be additive in the sense described above. So consider the part of New York City that's on Long Island. That's just Brooklyn plus Queens. If the population function is a measure, the value of that function, as applied to the Long Island part of New York, should be 2,500,000 plus 2,000,000, i.e. 4,500,000. And that makes sense: the population of Brooklyn plus Queens just is the population of Brooklyn plus the population of Queens.

Not every function from regions to numbers is a measure. Consider the function that takes a region of New York City as input, and returns as output the proportion of people in that region who are New York Mets fans. We can imagine that this function has the following values.

Borough	**Mets Proportion**
Brooklyn	0.6
Queens	0.75
Manhattan	0.5
The Bronx	0.25
Staten Island	0.5

Now think again about the part of New York we discussed above: the Brooklyn plus Queens part. What proportion of people in that part of the city are Mets fans? We certainly can't figure that out by just looking at the Brooklyn number from the above table, 0.6, and the Queens number, 0.75, and adding them together. That would yield the absurd result that the proportion of people in that part of the city who are Mets fans is 1.35.

That's to say, the function from a region to the proportion of people in that region who are Mets fans is *not* a measure. Measures are functions that are always additive over subregions. The value of the function applied to a whole region is the sum of the values the function takes when applied to the parts. 'Counting' functions, like population, have this property.

The measure function we looked at above takes real regions, parts of New York City, as inputs. But measures can also be defined over things that are suitably analogous to regions. Imagine a family of four children, named below, who eat the following amounts of meat at dinner.

Child	**Meat Consumption (g)**
Alice	400
Bruce	300
Chuck	200
Daria	100

We can imagine a function that takes a group of children (possibly including just one child, or even no children) as inputs, and has as output how many grams of meat those children ate. This function will be a measure. If the 'groups' contain just the one child, the values of the function will be given by the above table. If the group contains two children, the values will be given by the addition rule. So for the group consisting of Alice and Chuck, the value of the function will be 600. That's because the amount of meat eaten by Alice and Chuck just is the amount of meat eaten by Alice, plus the amount of meat eaten by Chuck. Whenever the value of a function, as applied to a group, is the sum of the values of the function as applied to the members, we have a measure function.

1.14 Normalised Measures

A measure function is defined over some regions. Usually one of those regions will be the 'universe' of the function; that is, the region made up of all those regions the function is defined over. In the case where the regions are regions of physical space, as in our New York example, that will just be the physical space consisting of all the smaller regions that are inputs to the function. In our New York example, the universe is just New York City. In cases where the regions are somewhat more metaphorical, as in the case of the children's meat-eating, the universe will also be defined somewhat more metaphorically. In that case, it is just the group consisting of the four children.

However the universe is defined, a normalised measure is simply a measure function where the value the function gives to the universe is 1. So for every sub-region of the universe, its measure can be understood as a proportion of the universe.

We can 'normalise' any measure by simply dividing each value through by the value of the universe. If we wanted to normalise our New York City population measure, we would simply divide all values by 7,500,000. The values we would then end up with are as follows.

Borough	**Population**
Brooklyn	1/3
Queens	4/15
Manhattan	1/5
The Bronx	2/15
Staten Island	1/3

Some measures may not have a well-defined universe, and in those cases we cannot normalise the measure. But generally normalisation is a simple matter of dividing everything by the value the function takes when applied to the whole universe. And the benefit of doing this is that it gives us a simple way of representing proportions.

1.15 Formalities

So far I've given a fairly informal description of what measures are, and what normalised measures are. In this section we're going to go over the details more formally. If you understand the concepts well enough already, or if you aren't familiar enough with set theory to follow this section entirely, you should feel free to skip forward to the next section. Note that this is a slightly simplified, and hence slightly inaccurate, presentation; we aren't focussing on issues to do with infinity.

A measure is a function m satisfying the following conditions.

1. The domain D is a set of sets.
2. The domain is closed under union, intersection and complementation with respect to the relevant universe U. That is, if $A \in D$ and $B \in D$, then $(A \cup B) \in D$ and $(A \cup B) \in D$ and $U \setminus A \in D$
3. The range is a set of non-negative real numbers
4. The function is additive in the following sense: If $A \cap B = \emptyset$, then $m(A \cup B) = m(A) + m(B)$

We can prove some important general results about measures using just these properties. Note that we the following results follow more or less immediately from additivity.

1. $m(A) = m(A \cap B) + m(A \cap (U \setminus B))$
2. $m(B) = m(A \cap B) + m(B \cap (U \setminus A))$
3. $m(A \cup B) = m(A \cap B) + m(A \cap (U \setminus B)) + m(B \cap (U \setminus A))$

The first says that the measure of A is the measure of A's intersection with B, plus the measure of A's intersection with the complement of B. The first says that the measure of B is the measure of A's intersection with B, plus the measure of B's intersection with the complement of A. In each case the point is that a set is just made up of its intersection with some other set, plus its intersection with the complement of that set. The final line relies on the fact that the union of A and B is made up of (i) their intersection, (ii) the part of A that overlaps B's complement and (iii) the part of B that overlaps A's complement. So the measure of $A \cup B$ should be the sum of the measure of those three sets.

Note that if we add up the LHS and RHS of lines 1 and 2 above, we get

$$m(A) + m(B) = m(A \cap B) + m(A \cap (U \setminus B)) + m(A \cap B) + m(A \cap (U \setminus B))$$

And subtracting $m(A \cap B)$ from each side, we get

$$m(A) + m(B) - m(A \cap B) = m(A \cap B) + m(A \cap (U \setminus B)) + m(A \cap (U \setminus B))$$

But that equation, plus line 3 above, entails that

$$m(A) + m(B) - m(A \cap B) = m(A \cup B)$$

And that identity holds whether or not $A \cap B$ is empty. If $A \cap B$ is empty, the result is just equivalent to the addition postulate, but in general it is a stronger result, and one we'll be using a fair bit in what follows.

1.16 Possibility Space

Imagine you're watching a baseball game. There are lots of ways we could get to the final result, but there are just two ways the game could end. The home team could win, call this possibility H, or the away team could win, call this possibility A.

Let's complicate the example somewhat. Imagine that you're watching one game while keeping track of what's going on in another game. Now there are four ways that the games could end. Both home teams could win. The home team could win at your game while the away team wins the other game. The away team could win at your game while the home team wins the other game. Or both away teams could win. This is a little easier to represent on a chart.

Your game	**Other game**
H	H
H	A
A	H
A	A

Here H stands for home team winning, and A stands for away team winning. If we start to consider a third game, there are now 8 possibilities. We started with 4 possibilities, but now each of these divides in 2: one where the home team wins the third game, and one where the away team wins. It's just about impossible to represent these verbally, so we'll just use a chart.

Game 1	**Game 2**	**Game 3**
H	H	H
H	H	A
H	A	H
H	A	A
A	H	H
A	H	A
A	A	H
A	A	A

Of course, in general we're interested in more things than just the results of baseball games. But the same structure can be applied to many more cases.

Say that there are three propositions, p, q and r that we're interested in. And assume that all we're interested in is whether each of these propositions is true or false. Then there are eight possible ways things could turn out, relative to what we're interested in. In the following table, each row is a possibility. T means the proposition at the head of that column is true, F means that it is false.

p	**q**	**r**
T	T	T
T	T	F
T	F	T
T	F	F
F	T	T
F	T	F
F	F	T
F	F	F

These eight possibilities are the foundation of the possibility space we'll use to build a probability function.

A measure is an additive function. So once you've set the values of the smallest parts, you've fixed the values of the whole. That's because for any larger part, you can work out its value by summing the values of its smaller parts. We can see this in the above example. Once you've fixed how much meat each child has eaten, you've fixed how much meat each group of children have eaten. The same goes for probability functions. In the cases we're interested in, once you've fixed the measure, i.e. the probability of each of the eight basic possibilities represented by the above eight rows, you've fixed the probability of all propositions that we're interested in.

For concreteness, let's say the probability of each row is given as follows.

p	**q**	**r**	
T	T	T	0.0008
T	T	F	0.008
T	F	T	0.08
T	F	F	0.8
F	T	T	0.0002
F	T	F	0.001
F	F	T	0.01
F	F	F	0.1

So the probability of the fourth row, where p is true while q and r are false, is 0.8. (Don't worry for now about where these numbers come from; we'll spend much more time on that in what follows.) Note that these numbers sum to 1. This is required; probabilities are **normalised** measures, so they must sum to 1.

Then the probability of any proposition is simply the sum of the probabilities of each row on which it is true. For instance, the probability of p is the sum of the probabilities of the first four rows. That is, it is $0.0008 + 0.008 + 0.08 + 0.8$, which is 0.8888.

To make more progress, we need to say a bit more about the relation between simple and complex sentences. The start of the next chapter may be very familiar to those who recently took introductory logic, but hopefully we will soon advance to less familiar material.

Chapter 2

Probability

2.1 Compound Sentences

Some sentences have other sentences as parts. We're going to be especially interested in sentences that have the following structures, where A and B are themselves sentences.

- A and B; which we'll write as $A \wedge B$
- A or B; which we'll write as $A \vee B$
- It is not the case that A; which we'll write as $\neg A$

What's special about these three compound formations is that the truth value of the whole sentence is fixed by the truth value of the parts. In fact, we can present the relationship between the truth value of the whole and the truth value of the parts using the truth tables discussed in the previous chapter. Here are the tables for the three connectives. First for and,

A	**B**	**A ∧ B**
T	T	T
T	F	F
F	T	F
F	F	F

Then for or. (Note that this is so-called *inclusive* disjunction. The whole sentence is true if both disjuncts are true.)

A	**B**	**A ∨ B**
T	T	T
T	F	T
F	T	T
F	F	F

Finally for not.

A	**¬A**
T	F
F	T

The important thing about this way of thinking about compound sentences is that it is *recursive*. I said above that some sentences have other sentences as parts. The easiest cases of this to think about are cases where A and B are atomic sentences, i.e. sentences that don't themselves have other sentences as parts. But nothing in the definitions we gave, or in the truth tables, requires that. A and B themselves could also be compound. And when they are, we can use truth tables to figure out how the truth value of the whole sentence relates to the truth value of its smallest constituents.

It will be easiest to see this if we work through an example. So let's spend some time considering the following sentence.

$$(p \wedge q) \vee \neg r$$

The sentence has the form $A \vee B$. But in this case A is the compound sentence $p \wedge q$, and B is the compound sentence $\neg r$. If we're looking at the possible truth values of the three sentences p, q and r, we saw in the previous chapter that there are 2^3, i.e. 8 possibilities. And they can be represented as follows.

p	**q**	**r**
T	T	T
T	T	F
T	F	T
T	F	F
F	T	T
F	T	F
F	F	T
F	F	F

It isn't too hard, given what we said above, to see what the truth values of $p \wedge q$, and of $\neg r$ will be in each of those possibilities. The first of these, $p \wedge q$, is true at a possibility just in case there's a T in the first column (i.e. p is true) and a T in the second column (i.e. q is true). The second sentence, $\neg r$ is true just in case there's an F in the third column (i.e. r is false). So let's represent all that on the table.

p	q	r	p ∧ q	¬r
T	T	T	T	F
T	T	F	T	T
T	F	T	F	F
T	F	F	F	T
F	T	T	F	F
F	T	F	F	T
F	F	T	F	F
F	F	F	F	T

Now the whole sentence is a disjunction, i.e. an or sentence, with the fourth and fifth columns representing the two disjuncts. So the whole sentence is true just in case either there's a T in the fourth column, i.e. $p \wedge q$ is true, or a T in the fifth column, i.e. $\neg r$ is true. We can represent that on the table as well.

p	q	r	p ∧ q	¬r	(p ∧ q) ∨ ¬r
T	T	T	T	F	T
T	T	F	T	T	T
T	F	T	F	F	F
T	F	F	F	T	T
F	T	T	F	F	F
F	T	F	F	T	T
F	F	T	F	F	F
F	F	F	F	T	T

And this gives us the full range of dependencies of the truth value of our whole sentence on the truth value of its parts.

This is relevant to probability because, as we've been stressing, probability is a measure over possibility space. So if you want to work out the probability of a sentence like $(p \wedge q) \vee \neg r$, one way is to work out the probability of each of the eight basic possibilities here, then work out at which of those possibilities $(p \wedge q) \vee \neg r$ is true, then sum the probabilities of those possibilities at which it is true. To illustrate this, let's again use the table of probabilities from the previous chapter.

p	**q**	**r**	
T	T	T	0.0008
T	T	F	0.008
T	F	T	0.08
T	F	F	0.8
F	T	T	0.0002
F	T	F	0.001
F	F	T	0.01
F	F	F	0.1

If those are the probabilities of each basic possibility, then the probability of $(p \wedge q) \vee \neg r$ is the sum of the values on the lines on which it is true. That is, it is the sum of the values on lines 1, 2, 4, 6 and 8. That is, it is 0.0008 + 0.008 + 0.8 + 0.001 + 0.1, which is 0.9098.

2.2 Equivalence, Entailment, Inconsistency, and Logical Truth

To a first approximation, we can define logical equivalence and logical entailment within the truth-table framework. The accounts we'll give here aren't quite accurate, and we'll make them a bit more precise in the next section. But they are on the right track, and they suggest some results that are, as it turns out, true in the more accurate structure.

If two sentences have the same pattern of Ts and Fs in their truth table, they are logically equivalent. Consider, for example, the sentences $\neg\mathbf{A} \vee \neg\mathbf{B}$ and $\neg(\mathbf{A} \wedge \mathbf{B})$. Their truth tables are given in the fifth and seventh columns of this table.

$\mathbf{A}$	$\mathbf{B}$	$\neg\mathbf{A}$	$\neg\mathbf{B}$	$\neg\mathbf{A} \vee \neg\mathbf{B}$	$\mathbf{A} \wedge \mathbf{B}$	$\neg(\mathbf{A} \wedge \mathbf{B})$
T	T	F	F	**F**	T	**F**
T	F	F	T	**T**	F	**T**
F	T	T	F	**T**	F	**T**
F	F	T	T	**T**	F	**T**

Note that those two columns are the same. That means that the two sentences are logically equivalent.

Now something important follows from the fact that the sentences are true in the same rows. For each sentence, the probability of the sentence is the sum of the probabilities of the rows in which it is true. But if the sentences are true in the same row, those are the same sums in each case. So the probability of the two sentences is the same. This leads to an important result.

- **Logically equivalent sentences have the same probability**

Note that we haven't quite proven this yet, because our account of logical equivalence is not quite accurate. But the result will turn out to hold when we fix that inaccuracy.

One of the notions that logicians care most about is *validity*. An argument with premises $A_1, A_2, ..., A_n$ and conclusion B is valid if it is impossible for the premises to be true and the conclusion false. Slightly more colloquially, if the premises are true, then the conclusion has to be true. Again, we can approximate this notion using truth tables. An argument is *invalid* if there is a line where the premises are true and the conclusion false. An argument is *valid* if there is no such line. That is, it is valid if in all possibilities where all the premises are true, the conclusion is also true.

When the argument that has A as its only premise, and B as its conclusion, is valid, we say that A **entails** B. If every line on the truth table where A is true is also a line where B is true, then A entails B.

Again, this has consequences for probability. The probability of a sentence is the sum of the probability of the possibilities in which it is true. If A entails B, then the possibilities where B is true will include all the possibilities where A is true, and may include some more. So the probability of B can't be *lower* than the probability of A. That's because each of these probabilities are sums of non-negative numbers, and each of the summands in the probability of A is also a summand in the probability of B.

- **If A entails B, then the probability of B is at least as great as the probability of A**

The argument we've given for this is a little rough, because we're working with an approximation of the definition of entailment, but it will turn out that the result goes through even when we tidy up the details.

Two sentences are **inconsistent** if they cannot be true together. Roughly, that means there is no line on the truth table where they are both true. Assume that A and B are inconsistent. So A is true at lines $L_1, L_2, ..., L_n$, and B is true at lines $L_{n+1}, ..., L_m$, where these do not overlap. So $A \vee B$ is true at lines

$L_1, L_2, ..., L_n, L_{n+1}, ..., L_m$. So the probability of A is the probability of L_1 plus the probability of L_2 plus ... plus the probability of L_n. And the probability of B is the probability of L_{n+1} plus ... plus the probability of L_m. And the probability of $A \vee B$ is the probability of L_1 plus the probability of L_2 plus ... plus the probability of L_n plus L_{n+1} plus ... plus the probability of L_m. That's to say

- **If A and B are inconsistent, then the probability of $A \vee B$ equals the probability of A plus the probability of B**

This is just the addition rule for measures transposed to probabilities. And it is a crucial rule, one that we will use all the time. (Indeed, it is sometimes taken to be the characteristic axiom of probability theory. We will look at axiomatic approaches to probability in a few sections time)

Finally, a **logical truth** is something that is true in virtue of logic alone. It is true in all possibilities, since what logic is does not change. A logical truth is entailed by any sentence. And a logical truth only entails other sentences.

Any sentence that is true in all possibilities must have probability 1. That's because probability is a *normalised* measure, and in a normalised measure, the measure of the universe is 1. And a logical truth is true at every point in the 'universe' of logical space.

- **Any logical truth has probability 1**

2.3 Two Important Results

None of the three connectives is particularly hard to process, but the rule for negation may well be the easiest of the lot. The truth value of $\neg A$ is just the opposite of the truth value of A. So if A is true at a line, then $\neg A$ is false. And if A is false at a line, then $\neg A$ is true. So exactly one of A and $\neg A$ is true at each line. So the sum of the probabilities of those propositions must be 1.

We can get to this result another way. It is easy to see that $A \vee \neg A$ is a logical truth by simply looking at its truth table.

A	**¬A**	**A ∨ ¬A**
T	F	T
F	T	T

The sentence $A \vee \neg A$ is true on each line, so it is a logical truth. And logical truths have probability 1. Now A and $\neg A$ are clearly inconsistent. So the probability of their disjunction equals the sum of their probabilities. That's to say, $\Pr(A \vee \neg A) = \Pr(A) + \Pr(\neg A)$. But $\Pr(A \vee \neg A) = 1$. So,

$$\Pr(A) + \Pr(\neg A) = 1$$

One important consequence of this is that the probabilities of A and $\neg A$ can't vary independently. Knowing how probable A is settles how probable $\neg A$ is.

The next result is slightly more complicated, but only a little. Consider the following table of truth values and probabilities.

Pr	**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

The variables in the first column represent the probability of each row. We can see from the table that the following results all hold.

1. $\Pr(A) = x_1 + x_2$, since A is true on the first and second lines
2. $\Pr(B) = x_1 + x_3$, since B is true on the first and third lines
3. $\Pr(A \wedge B) = x_1$, since $A \wedge B$ is true on the first line
4. $\Pr(A \vee B) = x_1 + x_2 + x_3$, since $A \vee B$ is true on the first, second and third lines

Adding the first and second lines together, we get

$$\Pr(A) + \Pr(B) = x_1 + x_2 + x_1 + x_3$$

And adding the third and fourth lines together, we get

$$\Pr(A \wedge B) + \Pr(A \vee B) = x_1 + x_1 + x_2 + x_3$$

And simply rearranging the variables a little reveals that

$$\Pr(A) + \Pr(B) = \Pr(A \wedge B) + \Pr(A \vee B)$$

Again, this is a result that we will use a lot in what follows.

2.4 Axioms of Probability

We've introduced probability so far through the truth tables. If you are concerned with some finite number, say n of sentences, you can make up a truth table with 2^n rows representing all the possible combinations of truth values for those sentences. And then a probability function is simply a measure defined over sets of those rows, i.e. sets of possibilities.

But we can also introduce probability more directly. A probability function is a function that takes sentences as inputs, has outputs in $[0, 1]$, and satisfies the following constraints.

- If A is a logical truth, then $\Pr(A) = 1$
- If A and B are logically equivalent, then $\Pr(A) = \Pr(B)$
- If A and B are logically disjoint, i.e. $\neg(A \wedge B)$ is a logical truth, then $\Pr(A) + \Pr(B) = \Pr(A \vee B)$

To get a feel for how these axioms operate, I'll run through a few proofs using the axioms. The results we prove will be familiar from the previous chapter, but the interest here is in seeing how the axioms interact with the definitions of logical truth, logical equivalence and logical disjointedness to derive familiar results.

- $\Pr(A) + \Pr(\neg A) = 1$

Proof: It is a logical truth that $A \vee \neg A$. This can be easily seen on a truth table. So by axiom 1, $\Pr(A \vee \neg A) = 1$. The truth tables can also be used to show that $\neg(A \wedge A)$ is a logical truth, so A and $\neg A$ are disjoint. So $\Pr(A) + \Pr(\neg A) = \Pr(A \vee \neg A)$. But since $\Pr(A \vee \neg A) = 1$, it follows that $\Pr(A) + \Pr(\neg A) = 1$.

- **If A is a logical falsehood, i.e. $\neg A$ is a logical truth, then** $\Pr(A) = 0$

Proof: If $\neg A$ is a logical truth, then by axiom 1, $\Pr(\neg A) = 1$. We just proved that $\Pr(A) + \Pr(\neg A) = 1$. From this it follows that $\Pr(A) = 0$.

- $\Pr(A) + \Pr(B) = \Pr(A \vee B) + \Pr(A \wedge B)$

Proof: First, note that A is logically equivalent to $(A \wedge B) \vee (A \wedge \neg B)$, and that $(A \wedge B)$ and $(A \wedge \neg B)$ are logically disjoint. We can see both these facts in the following truth table.

$\mathbf{A}$	$\mathbf{B}$	$\neg\mathbf{B}$	$(\mathbf{A}\wedge\mathbf{B})$	$(\mathbf{A}\wedge\neg\mathbf{B})$	$(\mathbf{A}\wedge\mathbf{B})\vee(\mathbf{A}\wedge\neg\mathbf{B})$
T	T	F	T	F	T
T	F	T	F	T	T
F	T	F	F	F	F
F	F	T	F	F	F

The first and sixth columns are identical, so A and $(A \wedge B) \vee (A \wedge \neg B)$. By axiom 2, that means that $\Pr(A) = \Pr((A \wedge B) \vee (A \wedge \neg B))$.

The fourth and fifth column never have a T on the same row, so $(A \wedge B)$ and $(A \wedge \neg B)$ are disjoint. That means that $\Pr((A \wedge B) \vee (A \wedge \neg B) = \Pr((A \wedge B) + \Pr(A \wedge \neg B)$. Putting the two results together, we get that $\Pr(A) = \Pr((A \wedge B) + \Pr(A \wedge \neg B)$.

The next truth table is designed to get us two results. First, that $A \vee B$ is equivalent to $B \vee (A \wedge \neg B)$. And second that B and $(A \wedge \neg B)$ are disjoint.

$\mathbf{A}$	$\mathbf{B}$	$\mathbf{A}\vee\mathbf{B}$	$\neg\mathbf{B}$	$\mathbf{A}\wedge\neg\mathbf{B}$	$\mathbf{B}\vee(\mathbf{A}\wedge\neg\mathbf{B})$
T	T	T	F	F	T
T	F	T	T	T	T
F	T	T	F	F	T
F	F	F	T	F	F

Note that the third column, $A \vee B$, and the sixth column, $B \vee (A \wedge \neg B)$, are identical. So those two propositions are equivalent. So $\Pr(A \vee B) = \Pr(B \vee (A \wedge \neg B))$.

Note also that the second column, B and the fifth column, $A \wedge \neg B$, have no Ts in common. So they are disjoint. So $\Pr(B \vee (A \wedge \neg B)) = \Pr(B) + \Pr(A \wedge \neg B)$. Putting the last two results together, we get that $\Pr(A \vee B) = \Pr(B) + \Pr(A \wedge \neg B)$.

If we add $\Pr(A \wedge B)$ to both sides of that last equation, we get $\Pr(A \vee B) + \Pr(A \wedge B) = \Pr(B) + \Pr(A \wedge \neg B) + \Pr(A \wedge B)$. But note that we already proved that $\Pr(A \wedge \neg B) + \Pr(A \wedge B) = \Pr(A)$. So we can rewrite $\Pr(A \vee B) + \Pr(A \wedge B) = \Pr(B) + \Pr(A \wedge \neg B) + \Pr(A \wedge B)$ as $\Pr(A \vee B) + \Pr(A \wedge B) = \Pr(B) + \Pr(A)$. And simply rearranging terms around gives us $\Pr(A) + \Pr(B) = \Pr(A \vee B) + \Pr(A \wedge B)$, which is what we set out to prove.

2.5 Truth Tables and Possibilities

So far we've been assuming that whenever we are interested in n sentences, there are 2^n possibilities. But this isn't always the case. Sometimes a combination of truth values doesn't express a real possibility. Consider, for example, the case where A = *Many people enjoyed the play*, and B = *Some people enjoyed the play*. Now we might start trying to draw up a truth table as follows.

A	**B**
T	T
T	F
F	T
F	F

But there's something deeply wrong with this table. The second line doesn't represent a real possibility. It isn't possible that it's true that many people enjoyed the play, but false that some people enjoyed the play. In fact there are only three real possibilities here. First, many people (and hence some people) enjoyed the play. Second, some people, but not many people, enjoyed the play. Third, no one enjoyed the play. That's all the possibilities that there are. There isn't a fourth possibility.

In this case, *A* entails *B*, which is why there is no possibility where *A* is true and *B* is false. In other cases there might be more complicated interrelations between sentences that account for some of the lines not representing real possibilities. Consider, for instance, the following case.

- *A* = Alice is taller than Betty
- *B* = Betty is taller than Carla
- *C* = Carla is taller than Alice

Again, we might try and have a regular, 8 line, truth table for these, as below.

A	**B**	**C**
T	T	T
T	T	F
T	F	T
T	F	F
F	T	T
F	T	F
F	F	T
F	F	F

But here the first line is not a genuine possibility. If Alice is taller than Betty, and Betty is taller than Carla, then Carla can't be taller than Alice. So there are, at most, 7 real possibilities here. (We'll leave the question of whether there are fewer than 7 possibilities as an exercise.) Again, one of the apparent possibilities is not real.

The chance that there are lines on the truth tables that don't represent real possibilities means that we have to modify several of the definitions we offered above. More carefully, we should say.

- Two sentences are A and B are logically equivalent if (and only if) they have the same truth value at every line on the truth table *that represents a real possibility*.
- Some sentences $A_1, ..., A_n$ **entail** a sentence B if (and only if) at every line which (a) represents a real possibility and (b) each of $A_1, ..., A_n$ is true, B is true. Another way of putting this is that the argument from $A_1, ..., A_n$ to B is **valid**.
- Two sentences A and B are logically disjoint if (and only if) there is no line which (a) represents a real possibility and (b) they are both true at that line

Surprisingly perhaps, we don't have to change the definition of a probability function all that much. We started off by saying that you got a probability function, defined over $A_1, ..., A_n$ by starting with the truth table for those sentences, all 2^n rows of it, and assigning numbers to each row in a way that they added up to 1. The probability of any sentence was then the sum of the numbers assigned to each row at which it is true.

This needs to be changed a little. If something does not represent a real possibility, then its negation is a logical truth. And all logical truths have to get probability 1. So we have to assign 0 to every row that does not represent a real possibility.

But that's the only change we have to make. Still, any way of assigning numbers to rows such that the numbers sum to 1, and any row that does not represent a real possibility is assigned 0, will be a probability function. And, as long as we are only interested in sentences with A_1, A_n as parts, any probability function can be generated this way.

So in fact all of the proofs in the previous chapter of the notes will still go through. There we generated a lot of results from the assumption that any probability function is a measure over the possibility space generated by a truth table. And that assumption is, strictly speaking, true. Any probability function is a measure over the possibility space generated by a truth table. It's true that some such measures are not probability functions because they assign positive values to lines that don't represent real possibilities. But that doesn't matter for the proofs we were making there.

The upshot is that we can, for the purposes of decision theory, continue to think about probability functions using truth tables. Occasionally we will have to be a little more careful, but for the most part, just assigning numbers to rows gives us all the basic probability theory we will need.

2.6 Propositions and Possibilities

There are many things we can be uncertain about. Some of these concern matters of fact, especially facts about the future. We can be uncertain about horseraces, or elections, or the weather. And some of them concern matters to do with mathematics or logic. We might be uncertain about whether two propositions are logically equivalent. Or we might be uncertain whether a particular mathematical conjecture is true or false.

Sometimes our uncertainty about a subject matter relates to both things. I'm writing this in the middle of hurricane season, and we're frequently uncertain about what the hurricanes will do. There are computer models to predict them, but the models are very complicated, and take hours to produce results even once all the data is in. So we might also be uncertain about a purely mathematical fact, namely what this model will predict given these inputs.

One of the consequences of the axioms for probability theory we gave above is that any logical truth, and for current purposes at least mathematical truths count as logical truths, get probability 1. This might seem counterintuitive. Surely we can sensibly say that such and such a mathematical claim is likely to be true, or probable to be true. Or we can say that someone's logical conjecture is probably false. How could it be that the axioms of probability say otherwise?

Well, the important thing to remember here is that what we're developing is a formal, mathematical notion. It remains an open question, indeed a deep philosophical question, whether that mathematical notion is useful in making sense of our intuitive, informal notion of what's more or less likely, or more or less probable. It is natural to think at this point that probability theory, the mathematical version, will not be of much help in modelling our uncertainty about logic or mathematics.

At one level this should not be too surprising. In order to use a logical/mathematical model, we have to use logic and mathematics. And to use logic and mathematics, we have to presuppose that they are given and available to use. But that's very close already to presupposing that they aren't at all uncertain. Now this little argument isn't very formal, and it certainly isn't meant to be a conclusive proof that there couldn't be a mathematical model of uncertainty about mathematics. But it's a reason to think that such a model would have to solve some tricky conceptual questions that a model of uncertainty about the facts does not have to solve.

And not only should this not be surprising, it should not necessarily be too worrying. In decision theory, what we're usually concerned with is uncertainty about the facts. It's possible that probability theory can be the foundation for an excellent model for uncertainty about the facts even if such a model is a terrible tool for understanding uncertainty about mathematics. In most areas of science, we don't expect every model to solve every problem. I mentioned above that at this time of year, we spend a lot of time looking at computer models of hurricane behaviour. Those models are not particularly useful guides to, say, snowfall over winter. (Let alone guides to who will win the next election.) But that doesn't make them bad hurricane models.

The same thing is going to happen here. We're going to try to develop a mathematical model for uncertainty about matters of fact. That model will be extremely useful, when applied to its intended questions. If you apply the model to uncertainty about mathematics, you'll get the crazy result that no mathematical question could ever be uncertain, because every mathematical truth gets probability 1, and every falsehood probability 0. That's not a sign the model is failing; it is a sign that it is being misapplied. (Caveat: Given that the model has limits, we might worry about whether its limits are being breached in some applications. This is a serious question about some applications of decision theory.)

To end this section, I want to note a connection between this section and two large philosophical debates. The first is about the relationship between mathematics and logic. The second is about the nature of propositions. I'll spend one all-too-brief paragraph on each.

I've freely moved between talk of logical truths and mathematical truths in the above. Whether this is appropriate turns out to be a tricky philosophical question. One view about the nature of mathematics, called logicisim, holds that mathematics is, in some sense, part of logic. If that's right, then mathematical truths are logical truths, and everything I've said is fine. But logicism is very controversial, to put it mildly. So we shouldn't simply assume that mathematical truths are logical truths. But we can safely assume the following disjunction is true. Either (a) simple arithmetical truths (which is all we've been relying on) are part of logic, or (b) the definition of a probability function needs to be clarified so all logical and (simple) mathematical truths get probability 1. With that assumption, everything I've said here will go through.

I've taken probability functions to be defined over sentences. But it is more common in mathematics, and perhaps more elegant, to define probability functions over sets of possibilities. Now some philosophers, most notably Robert Stalnaker, have argued that sets of possibilities also have a central philosophical role. They've argued that propositions, the things we believe, assert, are uncertain about etc, just are sets of possibilities. If that's right, there's a nice connection between the mathematical models of probability, and the psychological notion of uncertainty we're interested in. But this view is controversial. Many philosophers think that, especially in logic and mathematics, there are many distinct propositions that are true in the same possibilities. (One can be uncertain about one mathematical truth while being certain that another is true, they think.) In any case, one of the upshots of the discussion above is that we're going to write as if Stalnaker was right, i.e. as if sets of possibilities are the things that we are certain/uncertain about. We'll leave the tricky philosophical questions about whether he's actually right for another day.

2.7 Conditional Probability

So far we've talked simply about the probability of various propositions. But sometimes we're not interested in the absolute probability of a proposition, we're interested in its **conditional** probability. That is, we're interested in the probability of the proposition *assuming* or *conditional on* some other proposition obtaining.

For example, imagine we're trying to decide whether to go to a party. At first glance, we might think that one of the factors that is relevant to our decision is the probability that it will be a successful party. But on second thought that isn't particularly relevant at all. If the party is going to be unpleasant if we are there (because we'll annoy the host) but quite successful if we aren't there, then it might be quite probable that it will be a successful party, but that will be no reason at all for us to go. What matters is the probabiilty of it being a good, happy party *conditional* on our being there.

It isn't too hard to visualise how conditional probability works if we think of measures over lines on the truth table. If we assume that something , call it B is true, then we should 'zero out', i.e. assign probability 0, to all the possibilities where B doesn't obtain. We're now left with a measure over only the B-possibilities. The problem is that it isn't a normalised measure. The values will only sum to $\Pr(B)$, not to 1. We need to renormalise. So we divide by $\Pr(B)$ and we get a probability back. In a formula, we're left with

$$\Pr(A|B) = \frac{\Pr(A \wedge B)}{\Pr(B)}$$

We can work through an example of this using a table that we've seen once or twice in the past.

p	**q**	**r**	
T	T	T	0.0008
T	T	F	0.008
T	F	T	0.08
T	F	F	0.8
F	T	T	0.0002
F	T	F	0.001
F	F	T	0.01
F	F	F	0.1

Assume now that we're trying to find the conditional probability of p given q. We could do this in two different ways.

First, we could set the probability of any line where q is false to 0. So we will get the following table.

p	**q**	**r**	
T	T	T	0.0008
T	T	F	0.008
T	F	T	0
T	F	F	0
F	T	T	0.0002
F	T	F	0.001
F	F	T	0
F	F	F	0

The numbers don't sum to 1 any more. They sum to 0.01. So we need to divide everything by 0.01. It's sometimes easier to conceptualise this as multiplying by $1/\Pr(q)$, i.e. by multiplying by 100. Then we'll end up with:

p	**q**	**r**	
T	T	T	0.08
T	T	F	0.8
T	F	T	0
T	F	F	0
F	T	T	0.02
F	T	F	0.1
F	F	T	0
F	F	F	0

And since p is true on the top two lines, the 'new' probability of p is 0.88. That is, the conditional probability of p given q is 0.88. As we were writing things above, $\Pr(p|q) = 0.88$.

Alternatively we could just use the formula given above. Just adding up rows gives us the following numbers.

$$\begin{aligned}\Pr(p \wedge q) &= 0.0008 + 0.008 = 0.0088 \\ \Pr(q) &= 0.0008 + 0.008 + 0.0002 + 0.001 = 0.01\end{aligned}$$

Then we can apply the formula.

$$\begin{aligned}\Pr(p|q) &= \frac{\Pr(p \wedge q)}{\Pr(q)} \\ &= \frac{0.0088}{0.01} \\ &= 0.88\end{aligned}$$

2.8 Bayes Theorem

It is often easier to calculate conditional probabilities in the 'inverse' direction to what we are interested in. That is, if we want to know $\Pr(A|B)$, it might be much easier to discover $\Pr(B|A)$. In these cases, we use Bayes Theorem to get the right result. I'll state Bayes Theorem in two distinct ways, then show that the two ways are ultimately equivalent.

$$\begin{aligned}\Pr(A|B) &= \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \\ &= \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|\neg A)\Pr(\neg A)}\end{aligned}$$

These are equivalent because $\Pr(B) = \Pr(B|A)\Pr(A) + \Pr(B|\neg A)\Pr(\neg A)$. Since this is an independently interesting result, it's worth going through the proof of it. First note that

$$\begin{aligned}\Pr(B|A)\Pr(A) &= \frac{\Pr(A \wedge B)}{\Pr(A)}\Pr(A) \\ &= \Pr(A \wedge B)\end{aligned}$$

$$\begin{aligned}\Pr(B|\neg A)\Pr(\neg A) &= \frac{\Pr(\neg A \wedge B)}{Pr\neg(A)}\Pr(\neg A) \\ &= \Pr(\neg A \wedge B)\end{aligned}$$

Adding those two together we get

$$\begin{aligned}
\Pr(B|A)\Pr(A) + \Pr(B|\neg A)\Pr(\neg A) &= \Pr(A \wedge B) + \Pr(\neg A \wedge B) \\
&= \Pr((A \wedge B) \vee (\neg A \wedge B)) \\
&= \Pr(B)
\end{aligned}$$

The second line uses the fact that $A \wedge B$ and $\neg A \wedge B$ are inconsistent, which can be verified using the truth tables. And the third line uses the fact that $(A \wedge B) \vee (\neg A \wedge B)$ is equivalent to A, which can also be verified using truth tables. So we get a nice result, one that we'll have occasion to use a bit in what follows.

$$\Pr(B) = \Pr(B|A)\Pr(A) + \Pr(B|\neg A)\Pr(\neg A)$$

So the two forms of Bayes Theorem are the same. We'll often find ourselves in a position to use the second form.

One kind of case where we have occasion to use Bayes Theorem is when we want to know how significant a test finding is. So imagine we're trying to decide whether the patient has disease D, and we're interested in how probable it is that the patient has the disease conditional on them returning a test that's positive for the disease. We also know the following background facts.

- In the relevant demographic group, 5% of patients have the disease.
- When a patient has the disease, the test returns a position result 80% of the time
- When a patient does not have the disease, the test returns a negative result 90% of the time

So in some sense, the test is fairly reliable. It usually returns a positive result when applied to disease carriers. And it usually returns a negative result when applied to non-carriers. But as we'll see when we apply Bayes Theorem, it is very unreliable in another sense. So let A be that the patient has the disease, and B be that the patient returns a positive test. We can use the above data to generate some 'prior' probabilities, i.e. probabilities that we use prior to getting information about the test.

- $\Pr(A) = 0.05$, and hence $\Pr(\neg A) = 0.95$
- $\Pr(B|A) = 0.8$
- $\Pr(B|\neg A) = 0.1$

Now we can apply Bayes theorem in its second form.

$$\begin{aligned}
\Pr(A|B) &= \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A)+\Pr(B|\neg A)\Pr(\neg A)} \\
&= \frac{0.8 \times 0.05}{0.08 \times 0.05 + 0.1 \times 0.95} \\
&= \frac{0.04}{0.04 + 0.095} \\
&= \frac{0.04}{0.135} \\
&\approx 0.296
\end{aligned}$$

So in fact the probability of having the disease, conditional on having a positive test, is less than 0.3. So in that sense the test is quite unreliable.

This is actually a quite important point. The fact that the probability of B given A is quite high does not mean that the probability of A given B is equally high. By tweaking the percentages in the example I gave you, you can come up with cases where the probability of B given A is arbitrarily high, even 1, while the probability of A given B is arbitrarily low.

Confusing these two conditional probabilities is sometimes referred to as the *prosecutors' fallacy*, though it's not clear how many actual prosecutors are guilty of it! The thought is that some prosecutors start with the premise that the probability of the defendant's blood (or DNA or whatever) matching the blood at the crime scene, conditional on the defendant being innocent, is 1 in a billion (or whatever it exactly is). They conclude that the probability of the defendant being innocent, conditional on their blood matching the crime scene, is about 1 in a billion. Because of derivations like the one we just saw, that is a clearly invalid move.

2.9 Conditionalisation

The following two concepts seem fairly closely related.

- The probability of some hypothesis H given evidence E
- The new probability of hypothesis H when evidence E comes in

In fact these are distinct concepts, though there are interesting philosophical questions about how intimately they are connected.

The first one is a *static* concept. It says, at one particular time, what the probability of H is given E. It doesn't say anything about whether or not E actually obtains. It doesn't say anything about changing your views, or your probabilities. It just tells us something about our current probabilities, i.e. our current measure on possibility space. And what it tells us is what proportion of the space where E obtains is occupied by possibilities where H obtains. (The talk of 'proportion' here is potentially misleading, since there's no physical space to measure. What we care about is the measure of the $E \wedge H$ space as a proportion of the measure of the E space.)

The second one is a *dynamic* concept. It says what we do when evidence E actually comes in. Once this happens, old probabilities go out the window, because we have to adjust to the new evidence that we have to hand. If E indicates H, then the probability of H should presumably go up, for instance.

Because these are two distinct concepts, we'll have two different symbols for them. We'll use $\Pr(H|E)$ for the static concept, and $Pr_E(H)$ for the dynamic concept. So $\Pr(H|E)$ is what the current probability of H given E is, and $Pr_E(H)$ is what the probability of H will be when we get evidence E.

Many philosophers think that these two should go together. More precisely, they think that a rational agent always updates by *conditionalisation*. That's just to say that for any rational agent, $\Pr(H|E) = Pr_E(H)$. When we get evidence E, we always replace the probability of H with the probability of H given E.

The conditionalisation thesis occupies a quirky place in contemporary philosophy. On the one hand it is almost universally accepted, and an extremely interesting set of theoretical results have been built up using the assumption it is true. (Pretty much everything in Bayesian philosophy of science relies in one way or another on the assumption that conditionalisation is correct. And since Bayesian philosophy of science is a thriving research program, this is a non-trivial fact.) On the other hand, there are remarkably few direct, and plausible, arguments in favor of conditionalisation. In the absence of a direct argument we can say two things.

First, the fact that a lot of philosophers (and statisticians and economists etc) accept conditionalisation, and have derived many important results using it, is a reason to take it seriously. The research programs that are based around conditionalisation do not seem to be degenerating, or failing to produce new insights. Second, in a lot of everyday applications, conditionalisation seems to yield sensible results. The simplest cases here are cases involving card games or roulette wheels where we can specify the probabilities of various outcomes in advance.

Let's work through a very simple example to see this. A deck of cards has 52 cards, of which 13 are hearts. Imagine we're about to draw 2 cards, without replacement, from that deck, which has been well-shuffled. The probability that the first is a heart is $^{13}/_{52}$, or, more simply, $^{1}/_{4}$. If we assume that a heart has been taken out, e.g. if we draw a heart with the first card, the probability that we'll draw another heart if $^{12}/_{51}$. That is, conditional on the first card we draw being a heart, the probability that the second is a heart if 12/51.

Now imagine that we do actually draw the first card, and it's a heart. What should the probability be that the next card will be a heart? It seems like it should be 12/51. Indeed, it is hard to see what else it could be. If A is *The first card drawn is a heart* and B is *The second card drawn is a heart*, then it seems both $\Pr(A|B)$ and $Pr_B(A)$ should be 12/51. And examples like this could be multiplied endlessly.

The support here for conditionalisation is not just that we ended up with the same result. It's that we seem to be making the same calculations both times. In cases like this, when we're trying to figure out $\Pr(A|B)$, we pretend we're trying to work out $Pr_B(A)$, and then stop pretending when we've worked out the calculation. If that's always the right way to work out $\Pr(A|B)$, then $\Pr(A|B)$ should always turn out to be equal to $Pr_B(A)$. Now this argument goes by fairly quickly obviously, and we might want to look over more details before deriving very heavy duty results from the idea that updating is always by conditionalisation, but it's easy to see we might take conditionalisation to be a plausible model for updating probabilities.

2.10 Conglomerability

Here is a feature that we'd like an updating rule to have. If getting some evidence E will make a hypothesis H more probable, then not getting E will not also make H more probable. Indeed, in standard cases, not getting evidence that would have made H more probable should make H less probable. It would be very surprising if we could know, before running a test, that however it turns out some hypothesis H will be more probable at the end of the test than at the beginning of it. We might have to qualify this in odd cases where H is, e.g., that the test is completed. But in standard cases if H will be likely whether some evidence comes in or doesn't come in, then H should be already likely.

We'll say that an update rule is **conglomerable** if it has this feature, and **non-conglomerable** otherwise. That is, it is non-conglomerable iff there are H and E such that,

$$Pr_E(H) > \Pr(H) \text{and} Pr_{\neg E}(H) > \Pr(H)$$

Now a happy result for conditionalisation, the rule that says $P_E(H) = \Pr(H|E)$, is that it is conglomerable. This result is worth going over in some detail. Assume that $\Pr(H|E) > \Pr(H) \text{and} Pr_{\neg E}(H) > \Pr(H)$. Then we can derive a contradicton as follows

$$\begin{aligned}
\Pr(H) &= \Pr((H \wedge E) \vee (H \wedge \neg E)) && \text{since } H = (H \wedge E) \vee (H \wedge \neg E) \\
&= \Pr(H \wedge E) + Pr(H \wedge \neg E) \\
&\qquad \text{since } (H \wedge E) \text{ and } (H \wedge \neg E) \text{ are disjoint} \\
&= \Pr(H|E)\Pr(E) + \Pr(H|\neg E)\Pr(\neg E) \\
&\qquad \text{since } \Pr(H|E)\Pr(E) = \Pr(H \wedge E) \\
&> \Pr(H)\Pr(E) + \Pr(H)\Pr(\neg E) \\
&\qquad \text{since } \Pr(H|E) > \Pr(H) \text{ and } \Pr(H|\neg E) > \Pr(H) \\
&= \Pr(H)(\Pr(E) + \Pr(\neg E)) \\
&= \Pr(H)\Pr(E \vee \neg E) && \text{since } E \text{ and } \neg E \text{ are disjoint} \\
&= \Pr(H) && \text{since } \Pr(E \vee \neg E) = 1
\end{aligned}$$

Conglomerability is related to dominance. The dominance rule of decision making says (among other things) that if C_1 is preferable to C_2 given E, and C_1 is preferable to C_2 given $\neg E$, then C_1 is simply preferable to C_2. Conglomerability says (among other things) that if $\Pr(H)$ is greater than x given E, and it is greater than x given $\neg E$, then it is simply greater than x.

Contemporary decision theory makes deep and essential use of principles of this form, i.e. that if something holds given E, and given $\neg E$, then it simply holds. And one of the running themes of these notes will be sorting out just which such principles hold, and which do not hold. The above proof shows that we get one nice result relating conditional probability and simple probability which we can rely on.

2.11 Independence

The probability of some propositions depends on other propositions. The probability that I'll be happy on Monday morning is not independent of whether I win the lottery on the weekend. On the other hand, the probability that I win the lottery on the weekend is independent of whether it rains in Seattle next weekend. Formally, we define **probabilistic indepdendence** as follows.

- Propositions A and B are **independent** iff $\Pr(A|B) = \Pr(A)$.

There is something odd about this definition. We purported to define a relationship that holds between pairs of propositions. It looked like it should be a symmetric relation: A is independent from B iff B is independent from A. But the definition looks asymmetric: A and B play very different roles on the right-hand side of the definition. Happily, this is just an appearance. Assuming that A and B both have positive probability, we can show that $\Pr(A|B) = \Pr(A)$ is equivalent to $\Pr(B|A) = \Pr(B)$.

$$\begin{aligned}
&\Pr(A|B) = \Pr(A) \\
\Leftrightarrow\ &\frac{\Pr(A \wedge B)}{\Pr(B)} = \Pr(A) \\
\Leftrightarrow\ &\Pr(A \wedge B) = \Pr(A) \times \Pr(B) \\
\Leftrightarrow\ &\frac{\Pr(A \wedge B)}{\Pr(A)} = \Pr(B) \\
\Leftrightarrow\ &\Pr(B|A) = \Pr(B)
\end{aligned}$$

We've multiplied and divided by $\Pr(A)$ and $\Pr(B)$, so these equivalences don't hold if $\Pr(A)$ or $\Pr(B)$ is 0. But in other cases, it turns out that $\Pr(A|B) = \Pr(A)$ is equivalent to $\Pr(B|A) = \Pr(B)$. And each of these is equivalent to the claim that $\Pr(A \wedge B) = \Pr(A)\Pr(B)$. This is an important result, and one that we'll refer to a bit.

- For independent propositions, the probability of their conjunction is the product of their probabilities.
- That is, if A and B are independent, then $\Pr(A \wedge B) = \Pr(A)\Pr(B)$

This rule doesn't apply in cases where A and B are dependent. To take an extreme case, when A is equivalent to B, then $A \wedge B$ is equivalent to A. In that case, $\Pr(A \wedge B) = \Pr(A)$, not $\Pr(A)^2$. So we have to be careful applying this multiplication rule. But it is a powerful rule in those cases where it works.

2.12 Kinds of Independence

The formula $\Pr(A|B) = \Pr(A)$ is, by definition, what probabilistic independence amounts to. It's important to note that probabilistic dependence is very different from causal dependence, and so we'll spend a bit of time going over the differences.

The phrase 'causal dependence' is a little ambiguous, but one natural way to use it is that A causally depends on B just in case B causes A. If we use it that way, it is an *asymmetric* relation. If B causes A, then A doesn't cause B. But probabilistic dependence is *symmetric*. That's what we proved in the previous section.

Indeed, there will typically be a quite strong probabilistic dependence between effects and their causes. So not only is the probability that I'll be happy on Monday dependent on whether I win the lottery, the probability that I'll win the lottery is dependent on whether I'll be happy on Monday. It isn't causally dependent; my moods don't cause lottery results. But the probability of my winning (or, perhaps better, having won) is higher conditional on my being happy on Monday than on my not being happy.

One other frequent way in which we get probabilistic dependence without causal dependence is when we have common effects of a cause. So imagine that Fred and I jointly purchased some lottery tickets. If one of those tickets wins, that will cause each of us to be happy. So if I'm happy, that is some evidence that I won the lottery, which is some evidence that Fred is happy. So there is a probabilistic connection between my being happy and Fred's being happy. This point is easier to appreciate if we work through an example numerically. Make each of the following assumptions.

- We have a 10% chance of winning the lottery, and hence a 90% chance of losing.
- If we win, it is certain that we'll be happy. The probability of either of us not being happy after winning is 0.
- If we lose, the probability that we'll be unhappy is 0.5.
- Moreover, if we lose, our happiness is completely independent of one another, so conditional on losing, the proposition that I'm happy is independent of the proposition that Fred's happy

So conditional on losing, each of the four possible outcomes have the same probability. Since these probabilities have to sum to 0.9, they're each equal to 0.225. So we can list the possible outcomes in a table. In this table *A* is winning the lottery, *B* is my being happy and *C* is Fred's being happy.

A	**B**	*C*	Pr
T	T	T	0.1
T	T	F	0
T	F	T	0
T	F	F	0
F	T	T	0.225
F	T	F	0.225
F	F	T	0.225
F	F	F	0.225

Adding up the various rows tells us that each of the following are true.

- $\Pr(B) = 0.1 + 0.225 + 0.225 = 0.55$
- $\Pr(C) = 0.1 + 0.225 + 0.225 = 0.55$
- $\Pr(B \wedge C) = 0.1 + 0.225 = 0.325$

From that it follows that $\Pr(B|C) = {}^{0.325}/_{0.55} \approx 0.59$. So $\Pr(B|C) > \Pr(B)$. So B and C are not independent. Conditionalising on C raises the probability of B because it raises the probability of one of the possible causes of C, and that cause is also a possible cause of B.

Often we know a lot more about probabilistic dependence than we know about causal connections and we have work to do to figure out the causal connections. It's very hard, especially in for example public health settings, to figure out what is a cause-effect pair, and what is the result of a common cause. One of the most important research programs in modern statistics is developing methods for solving just this problem. The details of those methods won't concern us here, but we'll just note that there's a big gap between probabilistic dependence and causal dependence.

On the other hand, it is usually safe to infer probabilistic dependence from causal dependence. If E is one of the (possible) causes of H, then usually E will change the probabilities of H. We can perhaps dimly imagine exceptions to this rule.

So imagine that a quarterback is trying to decide whether to run or pass on the final play of a football game. He decides to pass, and the pass is successful, and his team wins. Now as it happens, had he decided to run, the team would have had just as good a chance of winning, since their run game was exactly as likely to score as their pass game. It's not crazy to think in those circumstances that the decision to pass was among the causes of the win, but the win was probabilistically independent of the decision to pass. In general we can imagine cases where some event moves a process down one of two possible paths to success, and where the other path had just as good a chance of success. (Imagine a doctor deciding to operate in a certain way, a politician campaigning in one area rather than another, a storm moving a battle from one piece of land to another, or any number of such cases.) In these cases we might have causal dependence (though whether we do is a contentious issue in the metaphysics of causation) without probabilistic dependence.

But such cases are rare at best. It is a completely commonplace occurrence to have probabilistic dependence without clear lines of causal dependence. We have to have very delicately balanced states of the world in order to have causal dependence without probabilistic dependence, and in every day cases we can safely assume that such a situation is impossible without probabilistic connections.

2.13 Gamblers' Fallacy

If some events are independent, then the probability of one is independent of the probability of the others. So knowing the results of one event gives you no guidance, not even probabilistic guidance, into whether the other will happen.

These points may seem completely banal, but in fact they are very hard to fully incorporate into our daily lives. In particular, they are very hard to completely incorporate in cases where we are dealing with successive outcomes of a particular chance process, such as a dice roll or a coin flip. In those cases we know that the individual events are independent of one another. But it's very hard not to think that, after a long run of heads say, that the coin landing tails is 'due'.

This feeling is what is known as the *Gamblers' Fallacy*. It is the fallacy of thinking that, when events A and B are independent, that what happens in A can be a guide of some kind to event B.

One way of noting how hard a grip the Gamblers' Fallacy has over our thoughts is to try to simulate a random device such as a coin flip. As an exercise, imagine that you're writing down the results of a series of 100 coin flips. Don't actually flip the coin, just write down a sequence of 100 Hs (for Heads) and Ts (for Tails) that look like what you think a random series of coin flips will look like. I suspect that it won't look a lot like what an actual sequence does look like, in part because it is hard to avoid the Gamblers' Fallacy.

Occasionally people will talk about the Inverse Gamblers' Fallacy, but this is a much less clear notion. The worry would be someone inferring from the fact that the coin has landed heads a lot that it will probably land heads next time. Now sometimes, if we know that it is a fair coin for example, this will be just as fallacious as the Gamblers' Fallacy itself. But it isn't always a fallacy. Sometimes the fact that the coin lands heads a few times in a row is evidence that it isn't really a fair coin.

It's important to remember the gap between causal and probabilistic dependence here. In normal coin-tossing situations, it is a mistake to think that the earlier throws have a causal impact on the later throws. But there are many ways in which we can have probabilistic dependence without causal dependence. And in cases where the coin has been landing heads a suspiciously large number of times, it might be reasonable to think that there is a common cause of it landing heads in the past and in the future - namely that it's a biased coin! And when there's a common cause of two causally independent events, they may be probabilistically dependent. That's to say, the first event might change the probabilities of the second event. In those cases, it doesn't seem fallacious to think that various patterns will continue.

This does all depend on just how plausible it is that there is such a causal mechanism. It’s one thing to think, because the coin has landed heads ten times in a row, that it might be biased. There are many causal mechanisms that could explain that. It’s another thing to think, because the coin has alternated heads and tails for the last ten tosses that it will continue to do so in the future. It’s very hard, in normal circumstances, to see what could explain that. And thinking that patterns for which there’s no natural causal explanation will continue is probably a mistake.

Chapter 3

Utility

3.1 Expected Values

A **random variable** is simply a variable that takes different numerical values in different states. In other words, it is a function from possibilities to numbers. Typically, random variables are denoted by capital letters. So we might have a random variable X whose value is the age of the next president of the United States, and his or her inauguration. Or we might have a random variable that is the number of children you will have in your lifetime. Basically any mapping from possibilities to numbers can be a random variable.

It will be easier to work with a specific example, so let's imagine the following case. You've asked each of your friends who will win the big football game this weekend, and 9 said the home team will win, while 5 said the away team will win. (Let's assume draws are impossible to make the equations easier.) Then we can let X be a random variable measuring the number of your friends who correctly predicted the result of the game. The value X takes is

$$X = \begin{cases} 9, & \text{if the home team wins,} \\ 5, & \text{if the away team wins.} \end{cases}$$

Given a random variable X and a probability function Pr, we can work out the **expected value** of that random variable with respect to that probability function. Intuitively, the expected value of X is a weighted average of the possible values of X, where the weights are given by the probability (according to Pr) of each value coming about. More formally, we work out the expected value of X this way. For each case, we multiply the value of X in that case by the probability of the case obtaining. Then we sum the numbers we've got, and the result is the expected value of X. We'll write the expected value of X as $Exp(X)$. So if the probability that the home wins is 0.8, and the probability that the away team wins is 0.2, then

$$\begin{aligned} Exp(X) &= 9 \times 0.8 + 5 \times 0.2 \\ &= 7.2 + 1 \\ &= 8.2 \end{aligned}$$

There are a couple of things to note about this result. First, the expected value of X isn't in any sense the value that we expect X to take. Indeed, the expected value of X is not even a value that X could take. So we shouldn't think that "expected value" is a phrase we can understand by simply understanding the notion of expectation and of value. Rather, we should think of the expected value as a kind of average.

Indeed, thinking of the expected value as an average lets us relate it back to the common notion of expectation. If you repeated the situation here – where there's an 0.8 chance that 9 of your friends will be correct, and an 0.2 chance that 5 of your friends will be correct – very often, then you would expect that in the long run the number of friends who were correct on each occasion would average about 8.2. That is, the expected value of a random variable X is what you'd expect the *average* value of X to be if (perhaps per impossible) the underlying situation was repeated many many times.

3.2 Maximise Expected Utility Rule

The orthodox view in modern decision theory is that the right decision is the one that maximises the expected utility of your choice. Let's work through a few examples to see how this might work. Consider again the decision about whether to take a cheap airline or a more reliable airline, where the cheap airline is cheaper, but it performs badly in bad weather. In cases where the probability is that the plane won't run into difficulties, and you have much to gain by taking the cheaper ticket, and even if something goes wrong it won't go badly wrong, it seems that you should take the cheaper plane. Let's set up that situation in a table.

	Good weather	Bad weather
	Pr = 0.8	Pr = 0.2
Cheap Airline	10	0
Reliable Airline	6	5

We can work out the expected utility of each action fairly easily.

$$\begin{aligned} Exp(\text{Cheap Airline}) &= 0.8 \times 10 + 0.2 \times 0 \\ &= 8 + 0 \\ &= 8 \\ Exp(\text{Reliable Airline}) &= 0.8 \times 6 + 0.2 \times 5 \\ &= 4.8 + 1 \\ &= 5.8 \end{aligned}$$

So the cheap airline has an expected utility of 8, the reliable airline has an expected utility of 5.8. The cheap airline has a higher expected utility, so it is what you should take.

We'll now look at three changes to the example. Each change should intuitively change the correct decision, and we'll see that the maximise expected utility rule does change in each case. First, change the downside of getting the cheap airline so it is now more of a risk to take it.

	Good weather	Bad weather
	Pr = 0.8	Pr = 0.2
Cheap Airline	10	-20
Reliable Airline	6	5

Here are the new expected utility considerations.

$$\begin{aligned} Exp(\text{Cheap Airline}) &= 0.8 \times 10 + 0.2 \times -20 \\ &= 8 + (-4) \\ &= 4 \\ Exp(\text{Reliable Airline}) &= 0.8 \times 6 + 0.2 \times 5 \\ &= 4.8 + 1 \\ &= 5.8 \end{aligned}$$

Now the expected utility of catching the reliable airline is higher than the expected utility of catching the cheap airline. So it is better to catch the reliable airline.

Alternatively, we could lower the price of the reliable airline, so it is closer to the cheap airline, even if it isn't quite as cheap.

	Good weather	Bad weather
	$Pr = 0.8$	$Pr = 0.2$
Cheap Airline	10	0
Reliable Airline	9	8

Here are the revised expected utility considerations.

$$\begin{aligned} Exp(\text{Cheap Airline}) &= 0.8 \times 10 + 0.2 \times 0 \\ &= 8 + 0 \\ &= 8 \\ Exp(\text{Reliable Airline}) &= 0.8 \times 9 + 0.2 \times 8 \\ &= 7.2 + 1.6 \\ &= 8.8 \end{aligned}$$

And again this is enough to make the reliable airline the better choice.

Finally, we can go back to the original utility tables and simply increase the probability of bad weather.

	Good weather	Bad weather
	$Pr = 0.3$	$Pr = 0.7$
Cheap Airline	10	0
Reliable Airline	6	5

We can work out the expected utility of each action fairly easily.

$$\begin{aligned} Exp(\text{Cheap Airline}) &= 0.3 \times 10 + 0.7 \times 0 \\ &= 3 + 0 \\ &= 3 \\ Exp(\text{Reliable Airline}) &= 0.3 \times 6 + 0.7 \times 5 \\ &= 1.8 + 3.5 \\ &= 5.3 \end{aligned}$$

We've looked at four versions of the same case. In each case the ordering of the outcomes, from best to worst, was:

1. Cheap airline and good weather
2. Reliable airline and good weather
3. Reliable airline and bad weather
4. Cheap airline and bad weather

As we originally set up the case, the cheap airline was the better choice. But there were three ways to change this. First, we increased the possible loss from taking the cheap airline. (That is, we increased the gap between the third and fourth options.) Second, we decreased the gain from taking the cheap airline.

(That is, we decreased the gap between the first and second options.) Finally, we increased the risk of things going wrong, i.e. we increased the probability of the bad weather state. Any of these on their own was sufficient to change the recommendation that "Maximise Expected Utility" makes. And that's all to the good, since any of these things does seem like it should be sufficient to change what's best to do.

3.3 Structural Features

When using the "Maximise Expected Utility" rule we assign a number to each choice, and then pick the option with the highest number. Moreover, the number we assign is independent of the other options that are available. The number we assign to a choice depends on the utility of that choice in each state and the probability of the states. Any decision rule that works this way is guaranteed to have a number of interesting properties.

First, it is guaranteed to be **transitive**. That is, if it recommends A over B, and B over C, then it recommends A over C. To see this, let's write the expected utility of a choice A as $Exp(U(A))$. If A is chosen over B, then $Exp(U(A)) > Exp(U(B))$. And if B is chosen over C, then $Exp(U(B)) > Exp(U(C))$. Now $>$, defined over numbers, is transitive. That is, if $Exp(U(A)) > Exp(U(B))$ and $Exp(U(B)) > Exp(U(C))$, then $Exp(U(A)) > Exp(U(C))$. So the rule will recommend A over B.

Second, it satisfies the independence of irrelevant alternatives. Assume A is chosen over B and C. That is, $Exp(U(A)) > Exp(U(B))$ and $Exp(U(A)) > Exp(U(C))$. Then A will be chosen when the only options are A and B, since $Exp(U(A)) > Exp(U(B))$. And A will be chosen when the only options are A and C, since $Exp(U(A)) > Exp(U(C))$. These two features are intuitively pleasing features of a decision rule.

Numbers are totally ordered by $>$. That is, for any two numbers x and y, either $x > y$ or $y > x$ or $x = y$. So if each choice is associated with a number, a similar relation holds among choices. That is, either A is preferable to B, or B is preferable to A, or they are equally preferable.

Expected utility maximisation never recommends choosing dominated options. Assume that A dominates B. For each state S_i, write utility of A in S_i as $U(A|S_i)$. Then dominance means that for all i, $U(A|S_i) > U(B|S_i)$. Now $Exp(U(A))$ and $Exp(U(B))$ are given by the following formulae. (In what follows n is the number of possible states.)

$$Exp(A) = \Pr(S_1)U(A|S_1) + \Pr(S_2)U(A|S_2) + ... + \Pr(S_n)U(A|S_n)$$
$$Exp(B) = \Pr(S_1)U(B|S_1) + \Pr(S_2)U(B|S_2) + ... + \Pr(S_n)U(B|S_n)$$

Note that the two values are each the sum of n terms. Note also that, given dominance, each term on the top row is at least as great as than the term immediately below it on the second row. (This follows from the fact that $U(A|S_i) > U(B|S_i)$ and the fact that $\Pr(S_i) \geq 0$.) Moreover, at least one of the terms on the top row is greater than the term immediately below it. (This follows from the fact that $U(A|S_i) > U(B|S_i)$ and the fact that for at least one i, $\Pr(S_i) > 0$. That in turn has to be true because if $\Pr(S_i) = 0$ for each i, then $\Pr(S_1 \vee S_2 \vee ... \vee S_n) = 0$. But $S_1 \vee S_2 \vee ... \vee S_n$ has to be true.) So $Exp(A)$ has to be greater than $Exp(B)$. So if A dominates B, it has a higher expected utility.

3.4 Generalising Dominance

The maximise expected utility rule also supports a more general version of dominance. We'll state the version of dominance using an example, then spend some time going over how we know maximise expected utility satisfies that version.

The original dominance principle said that if A is better than B in every state, then A is simply better than B simply. But we don't have to just compare choices in individual states, we can also compare them across any number of states. So imagine that we have to choose between A and B and we know that one of four states obtains. The utility of each choice in each state is given as follows.

	S_1	S_2	S_3	S_4
A	10	9	9	0
B	8	3	3	3

And imagine we're using the maximin rule. Then the rule says that A does better than B in S_1, while B does better than A in S_4. The rule also says that B does better than A overall, since it's worst case scenario is 3, while A's worst case scenario is 0. But we can also compare A and B with respect to pairs of states. So conditional on us just being in S_1 or S_2, then A is better. Because between those two states, its worst case is 9, while B's worst case is 3.

Now imagine we've given up on maximin, and are applying a new rule we'll call maxiaverage. The maxiaverage rule tells us make the choice that has the highest (or **maxi**mum) average of best case and worst case scenarios. The rule says that B is better overall, since it has a best case of 8 and a worst case of 3 for an average of 5.5, while A has a best case of 10 and a worst case of 0, for an average of 5.

But if we just know we're in S_1 or S_2, then the rule recommends A over B. That's because among those two states, A has a maximum of 10 and a minimum of 9, for an average of 9.5, while B has a maximum of 8 and a minimum of 3 for an average of 5.5.

And if we just know we're in S_3 or S_4, then the rule also recommends A over B. That's because among those two states, A has a maximum of 9 and a minimum of 0, for an average of 4.5, while B has a maximum of 3 and a minimum of 3 for an average of 3.

This is a fairly odd result. We know that either we're in one of S_1 or S_2, or that we're in one of S_3 or S_4. And the rule tells us that if we find out which, i.e. if we find out we're in S_1 or S_2, or we find out we're in S_3 or S_4, either way we should choose A. But before we find this out, we should choose B.

Here then is a more general version of dominance. Assume our initial states are $\{S_1, S_2, ..., S_n\}$. Call this set S. A binary partition of S is a pair of sets of states, call them T_1 and T_2, such that every state in S is in exactly one of T_1 and T_2. (We're simplifying a little here - generally a partition is any way of dividing a collection up into parts such that every member of the original collection is in one of the 'parts'. But we'll only be interested in cases where we divide the original states in two, i.e., into a *binary* partition.) Then the generalised version of dominance says that if A is better than B among the states in T_1, and it is better than B among the states in T_2, where T_1 and T_2 provide a partition of S, then it is better than B among the states in S. That's the principle that maxiaverage violates. A is better than B among the states $\{S_1, S_2\}$. And it is better than B among the states $\{S_3, S_4\}$. But it isn't better than B among the states $\{S_1, S_2, S_3, S_4\}$. That is, it isn't better than B among the states generally.

We'll be interested in this principle of dominance because, unlike perhaps dominance itself, there are some cases where it leads to slightly counterintuitive results. For this reason some theorists have been interested in theories which, although they satisfy dominance, do not satisfy this general version of dominance.

On the other hand, maximise expected utility does respect this principle. In fact, it respects an even stronger principle, one that we'll state using the notion of **conditional expected utility**. Recall that as well as probabilities, we defined conditional probabilities above. Well conditional expected utilities are just the expectations of the utility function with respect to a conditional probability. More formally, if there are states $S_1, S_2, ..., S_n$, then the expected utility of A conditional on E, which we'll write $Exp(U(A|E)$, is

$$Exp(U(A|E)) = \Pr(S_1|E)U(S_1|A) + \Pr(S_2|E)U(S_2|A) + ... + \Pr(S_n|E)U(S_n|A)$$

That is, we just replace the probabilities in the definition of expected utility with conditional probabilities. (You might wonder why we didn't also replace the utilities with conditional utilities. That's because we're assuming that states are defined so that given an action, the state has a fixed utility. If we didn't make this simplifying assumption, we'd have to be more careful here.) Now we can prove the following theorem.

- If $Exp(U(A|E)) > Exp(U(B|E))$, and $Exp(U(B|\neg E)) > Exp(U(B|\neg E))$, then $Exp(U(A)) > Exp(U(B))$.

We'll prove this by proving something else that will be useful in many contexts.

- $Exp(U(A)) = Exp(U(A|E))\Pr(E) + Exp(U(A|\neg E))\Pr(\neg E)$

To see this, note the following

$$\begin{aligned} Pr(S_i) &= \Pr((S_i \wedge E) \vee (S_i \wedge \neg E)) \\ &= \Pr(S_i \wedge E) + \Pr(S_i \wedge \neg E) \\ &= \Pr(S_i|E)\Pr(E) + \Pr(S_i|\neg E)\Pr(\neg E) \end{aligned}$$

And now we'll use this when we're expanding $Exp(U(A|E))\Pr(E)$.

$$\begin{aligned} Exp(U(A|E))\Pr(E) &= \Pr(E)[Pr(S_1|E)U(S_1|A) + \Pr(S_2|E)U(S_2|A) \\ &\quad + ... + \Pr(S_n|E)U(S_n|A)] \\ &= \Pr(E)\Pr(S_1|E)U(S_1|A) + \Pr(E)\Pr(S_2|E)U(S_2|A) \\ &\quad + ... + \Pr(E)\Pr(S_n|E)U(S_n|A) \\ Exp(U(A|\neg E))\Pr(\neg E) &= \Pr(\neg E)[Pr(S_1|\neg E)U(S_1|A) + \Pr(S_2|\neg E)U(S_2|A) \\ &\quad + ... + \Pr(S_n|\neg E)U(S_n|A)] \\ &= \Pr(\neg E)\Pr(S_1|\neg E)U(S_1|A) \\ &\quad + \Pr(\neg E)\Pr(S_2 \neg |E)U(S_2|A) \\ &\quad + ... + \Pr(\neg E)\Pr(S_n|\neg E)U(S_n|A) \end{aligned}$$

Putting those two together, we get

$$\begin{aligned} & Exp(U(A|E))\Pr(E) + Exp(U(A|\neg E))\Pr(\neg E) \\ &\quad = \Pr(E)\Pr(S_1|E)U(S_1|A) + ... + \Pr(E)\Pr(S_n|E)U(S_n|A) + \\ &\qquad \Pr(\neg E)\Pr(S_1|\neg E)U(S_1|A) + ... + \Pr(\neg E)\Pr(S_n|\neg E)U(S_n|A) \\ &\quad = (\Pr(E)\Pr(S_1|E) + \Pr(\neg E)\Pr(S_1|\neg E))U(S_1|A) \\ &\qquad + ... + (\Pr(E)\Pr(S_n|E) + \Pr(\neg E)\Pr(S_n|\neg E))U(S_n|A) \\ &\quad = \Pr(S_1)U(S_1|A) + \Pr(S_2)U(S_2|A) + ... \Pr(S_n)U(S_n|A) \\ &\quad = Exp(U(A)) \end{aligned}$$

Now if $Exp(U(A|E)) > Exp(U(B|E))$, and $Exp(U(B|\neg E)) > Exp(U(B|\neg E))$, then the following two inequalities hold.

$$\begin{gathered} Exp(U(A|E))\Pr(E) \geq Exp(U(B|E))\Pr(E) \\ Exp(U(A|\neg E))\Pr(\neg E) \geq Exp(U(B|\neg E))\Pr(\neg E) \end{gathered}$$

In each case we have equality only if the probability in question ($Pr(E)$ in the first line, $Pr(\neg E)$ in the second) is zero. Since not both $Pr(E)$ and $Pr(\neg E)$ are zero, one of those is a strict inequality. (That is, the left hand side is greater than, not merely greater than or equal to, the right hand side.) So adding up the two lines, and using the fact that in one case we have a strict inequality, we get

$$\begin{aligned} & Exp(U(A|E))\Pr(E) + Exp(U(A|\neg E))\Pr(\neg E) > \\ & Exp(U(B|E))\Pr(E) + Exp(U(B|\neg E))\Pr(\neg E) \\ \text{i.e. } & Exp(U(A)) > Exp(U(B)) \end{aligned}$$

That is, if A is better than B conditional on E, and it is better than B conditional on $\neg E$, then it is simply better than B.

3.5 Sure Thing Principle

The result we just proved is very similar to a famous principle of decision theory, the Sure Thing Principle. The Sure Thing Principle is usually stated in terms of one option being at least as good as another, rather than one option being better than another, as follows.

Sure Thing Principle If $AE \succeq BE$ and $A\neg E \succeq B\neg E$, then $A \succeq B$.

The terminology there could use some spelling out. By $A \succ B$ we mean that A is preferred to B. By $A \succeq B$ we mean that A is regarded as at least as good as B. The relation between $\succ$ and $\succeq$ is like the relation between $>$ and $\geq$. In each case the line at the bottom means that we're allowing equality between the values on either side.

The odd thing here is using $AE \succeq BE$ rather than something that's explicitly conditional. We should read the terms on each side of the inequality sign as *conjunctions*. It means that A *and* E is regarded as at least as good an outcome as B and E. But that sounds like something that's true just in case the agent prefers A to B conditional on E obtaining. So we can use preferences over conjunctions like AE as proxy for conditional preferences.

So we can read the Sure Thing Principle as saying that if A is at least as good as B conditional on E, and conditional on $\neg E$, then it really is at least as good as

B. Again, this looks fairly plausible in the abstract, though we'll soon see some reasons to worry about it.

Expected Utility maximisation satisfies the Sure Thing Principle. I won't go over the proof here because it's really just the same as the proof from the previous section with $>$ replaced by $\geq$ in a lot of places. But if we regard the Sure Thing Principle as a plausible principle of decision making, then it is a good feature of Expected Utility maximisation that it satisfies it.

It is tempting to think of the Sure Thing Principle as a generalisation of a principle of logical implication we all learned in propositional logic. The principle in question said that from $X \rightarrow Z$, and $Y \rightarrow Z$, and $X \vee Y$, we can infer C. If we let Z be that A is better than B, let X be E, and Y be $\neg E$, it looks like we have all the premises, and the reasoning looks intuitively right. But this analogy is misleading for two reasons.

First, for technical reasons we can't get into in depth here, preferring A to B conditional on E isn't the same as it being true that if E is true you prefer A to B. To see some problems with this, think about cases where you don't know E is true, and A is something quite horrible that mitigates the effects of the unpleasant E. In this case you do prefer AE to BE, and E is true, but you don't prefer A to B. But we'll set this question, which is largely a logical question about the nature of conditionals, to one side.

The bigger problem is that the analogy with logic would suggest that the following generalisation of the Sure Thing Principle will hold.

Disjunction Principle If $AE_1 \succeq BE_1$ and $AE_2 \succeq BE_2$, and $Pr(E_1 \vee E_2) = 1$ then $A \succeq B$.

But this "Disjunction Principle" seems no good in cases like the following. I'm going to toss two coins. Let p be the proposition that they will land differently, i.e. one heads and one tails. I offer you a bet that pays you \$2 if p, and costs you \$3 if $\neg p$. This looks like a bad bet, since $Pr(p) = 0.5$, and losing \$3 is worse than gaining \$2. But consider the following argument.

Let E_1 be that at least one of the coins landing heads. It isn't too hard to show that $Pr(p|E_1) = 2/3$. So conditional on E_1, the expected return of the bet is $2/3 \times 2 - 1/3 \times 3 = 4/3 - 1 = 1/3$. That's a positive return. So if we let A be taking the bet, and B be declining the bet, then conditional on E_1, A is better than B, because the expected return is positive.

Let E_2 be that at least one of the coins landing tails. It isn't too hard to show that $Pr(p|E_1) = 2/3$. So conditional on E_2, the expected return of the bet is $2/3 \times 2 - 1/3 \times 3 = 4/3 - 1 = 1/3$. That's a positive return. So if we let A be taking the bet, and B be declining the bet, then conditional on E_2, A is better than B, because the expected return is positive.

Now if E_1 fails, then both of the coins lands tails. That means that at least one of the coins lands tails. That means that E_2 is true. So if $E1$ fails $E2$ is true. So one of $E1$ and $E2$ has to be true, i.e. $Pr(E_1 \vee E_2) = 1$. And $AE_1 \succeq BE_1$ and $AE_2 \succeq BE_2$. Indeed $AE_1 \succ BE_1$ and $AE_2 \succ BE_2$. But $B \succ A$. So the disjunction principle isn't in general true.

It's a deep philosophical question how seriously we should worry about this. If the Sure Thing Principle isn't any more plausible intuitively than the Disjunction Principle, and the Disjunction Principle seems false, does that mean we should be sceptical of the Sure Thing Principle? As I said, that's a very hard question, and it's one we'll return to a few times in what follows.

3.6 Allais Paradox

The Sure Thing Principle is one of the more controversial principles in decision theory because there seem to be cases where it gives the wrong answer. The most famous of these is the Allais paradox, first discovered by the French economist (and Nobel Laureate) Maurice Allais. In this paradox, the subject is first offered the following choice between A and B. The results of their choice will depend on the drawing of a coloured ball from an urn. The urn contains 10 white balls, 1 yellow ball, and 89 black balls, and assume the balls are all randomly distributed so the probability of drawing each is identical.

	White	Yellow	Black
A	\$1,000,000	\$1,000,000	\$0
B	\$5,000,000	\$0	\$0

That is, they are offered a choice between an 11% shot at \$1,000,000, and a 10% shot at \$5,000,000. Second, the subjects are offered the following choice between C and D, which are dependent on drawings from a similarly constructed urn.

	White	Yellow	Black
C	\$1,000,000	\$1,000,000	\$1,000,000
D	\$5,000,000	\$0	\$1,000,000

That is, they are offered a choice between \$1,000,000 for sure, and a complex bet that gives them a 10% shot at \$5,000,000, an 89% shot at \$1,000,000, and a 1% chance of striking out and getting nothing.

Now if we were trying to maximise expected *dollars*, then we'd have to choose both B and D. But, and this is an important point that we'll come back to, dollars aren't utilities. Getting \$2,000,000 isn't twice as good as getting \$1,000,000.

pretty clearly if you were offered a million dollars or a 50% chance at two million dollars you would, and should, take the million for sure. That's because the two million isn't twice as useful to you as the million. Without a way of figuring out the utility of \$1,000,000 versus the utility of \$5,000,000, we can't say whether A is better than B. But we can say one thing. You can't consistently hold the following three views.

- $B \succ A$
- $C \succ D$
- The Sure Thing Principle holds

This is relevant because a lot of people think $B \succ A$ and $C \succ D$. Let's work through the proof of this to finish with.

Let E be that either a white or yellow ball is drawn. So $\neg E$ is that a black ball is drawn. Now note that $A\neg E$ is identical to $B\neg E$. In either case you get nothing. So $A\neg E \succeq B\neg E$. So if $AE \succeq BE$ then, by Sure Thing, $A \succeq B$. Equivalently, if $B \succ A$, then $BE \succ AE$. Since we've assumed $B \succ A$, then $BE \succ AE$.

Also note that $C\neg E$ is identical to $D\neg E$. In either case you get a million dollars. So $D\neg E \succeq C\neg E$. So if $DE \succeq CE$ then, by Sure Thing, $D \succeq C$. Equivalently, if $C \succ D$, then $CE \succ DE$. Since we've assumed $C \succ D$, then $CE \succ DE$.

But now we have a problem, since $BE = DE$, and $AE = CE$. Given E, then choice between A and B just is the choice between C and D. So holding simultaneously that $BE \succ AE$ and $CE \succ DE$ is incoherent.

It's hard to say for sure just what's going on here. Part of what's going on is that we have a 'certainty premium'. We prefer options like C that guarantee a positive result. Now having a certainly good result is a kind of holistic property of C. The Sure Thing Principle in effect rules out assigning value to holistic properties like that. The value of the whole need not be *identical* to the value of the parts, but any comparisons between the values of the parts has to be reflected in the value of the whole. Some theorists have thought that a lesson of the Allais paradox is that this is a mistake.

We won't be looking in this course at theories which violate the Sure Thing Principle, but we will be looking at justifications of the Sure Thing Principle, so it is worth thinking about reasons you might have for rejecting it.

Chapter 4

Working Out Probabilities

As might be clear from the discussion of what probability functions are, there are a lot of probability functions. For instance, the following is a probability function for any (logically independent) p and q.

p	**q**	Pr
T	T	0.97
T	F	0.01
F	T	0.01
F	F	0.01

But if p actually is that the moon is made of green cheese, and q is that there are little green men on Mars, you probably won't want to use this probability function in decision making. That would commit you to making some bets that are intuitively quite crazy.

So we have to put some constraints on the kinds of probability we use if the "Maximise Expected Utility" rule is likely to make sense. As it is sometimes put, we need to have an **interpretation** of the Pr in the expected utility rule. We'll look at three possible interpretations that might be used.

4.1 Frequency

Historically probabilities were often identified with frequencies. If we say that the probability that this F is a G is, say, $\frac{2}{3}$, that means that the proportion of F's that are G's is $\frac{2}{3}$.

Such an approach is plausible in a lot of cases. If we want to know what the probability is that a particular student will catch influenza this winter, a good first step would be to find out the proportion of students who will catch influenza this winter. Let's say this is $\frac{1}{10}$. Then, to a first approximation, if we need to feed into our expected utility calculator the probability that this student will catch influenza this winter, using $\frac{1}{10}$ is not a bad first step. Indeed, the insurance industry does not a bad job using frequencies as guides to probabilities in just this way.

But that can hardly be the end of the story. If we know that this particular student has not had an influenza shot, and that their boyfriend and their roommate have both caught influenza, then the probability of them catching influenza would now be much higher. With that new information, you wouldn't want to take a bet that paid \$1 if they didn't catch influenza, but lost you \$8 if they did catch influenza. The odds now look like that's a bad bet.

Perhaps the thing to say is that the relevant group is not all students. Perhaps the relevant group is students who haven't had influenza shots and whose roommates and boyfriends have also caught influenza. And if, say, $\frac{2}{3}$ of such students have caught influenza, then perhaps the probability that this student will catch influenza is $\frac{2}{3}$.

You might be able to see where this story is going by now. We can always imagine more details that will make that number look inappropriate as well. Perhaps the student in question is spending most of the winter doing field work in South America, so they have little chance to catch influenza from their infected friends. And now the probability should be lower. Or perhaps we can imagine that they have a genetic predisposition to catch influenza, so the probability should be higher. There is always more information that could be relevant.

The problem for using frequencies as probabilities then is that there could always be more precise information that is relevant to the probability. Every time we find that the person in question isn't merely an F (a student, say), but is a particular kind of F (a student who hasn't had an influenza shot, whose close contacts are infected, who has a genetic predisposition to influenza), we want to know the proportion not of F's who are G's, but the proportion of the more narrowly defined class who are G's. But eventually this will leave us with no useful probabilities at all, because we'll have found a way of describing the student in question such that they are the only person in history who satisfies this description.

This is hardly a merely theoretical concern. If we are interested in the probability that a particular bank will go bankrupt, or that a particular Presidential candidate will win election, it isn't too hard to come up with a list of characteristics of the bank or candidate in question in such a way that they are the only one in history to meet that description. So the frequency that such banks will go bankrupt is either 1 (1 out of 1 go bankrupt) or 0 (0 out of 1 do). But those aren't particularly useful probabilities. So we should look elsewhere for an interpretation of the Pr that goes into our definition of expected utility.

In the literature there are two objections to using frequencies as probabilities that seem related to the argument we're looking at here.

One of these is the **Reference Class Problem**. This is the problem that if we're interested in the probability that a particular person is G, then the frequency of G-hood amongst the different classes the person is in might differ.

The other is the **Single Case Problem**. This is the problem that we're often interested in one-off events, like bank failures, elections, wars etc, that don't naturally fit into any natural broader category.

I think the reflections here support the idea that these are two sides of a serious problem for the view that probabilities are frequencies. In general, there actually is a natural solution to the Reference Class Problem. We look to the most narrowly drawn reference class we have available. So if we're interested in whether a particular person will survive for 30 years, and we know they are a 52 year old man who smokes, we want to look not to the survival frequencies of people in general, or men in general, or 52 year old men in general, but 52 year old male smokers.

Perhaps by looking at cases like this, we can convince ourselves that there is a natural solution to the Reference Class Problem. But the solution makes the Single Case Problem come about. Pretty much anything that we care about is distinct in some way or another. That's to say, if we look closely we'll find that the most natural reference class for it just contains that one thing. That's to say, it's a single case in some respect. And one-off events don't have interesting frequencies. So frequencies aren't what we should be looking to as probabilities.

4.2 Degrees of Belief

In response to these worries, a lot of philosophers and statisticians started thinking of probability in purely subjective terms. The probability of a proposition p is just how confident the agent is that p will obtain. This level of confidence is the agent's *degree of belief* that p will obtain.

Now it isn't altogether easy to measure degrees of belief. I might be fairly confident that my baseball team will win tonight, and more confident that they'll win at least one of the next three games, and less confident that they'll win all of their next three games, but how could we measure numerically each of those strengths? Remember that probabilities are *numbers*. So if we're going to identify probabilities with degrees of belief, we have to have a way to convert strengths of confidence to numbers.

The core idea about how to do this uses the very decision theory that we're looking for input to. I'll run through a rough version of how the measurement works; this should be enough to give you the idea for what is going on. Imagine you have a chance to buy a ticket that pays \$1 if p is true. How much, in dollars, is the most would you pay for this? Well, it seems that how much you should pay for this is the probability of p. Let's see why this is true. (Assume in what follows that the utility of each action is given by how many dollars you get from the action; this is the simplifying assumption we're making.) If you pay \$Pr(p) for the ticket, then you've performed some action (call it A) that has the following payout structure.

$$U(A) = \begin{cases} 1 - \Pr(p) & \text{if } p, \\ -\Pr(p) & \text{if } \neg p. \end{cases}$$

So the expected value of $U(A)$ is

$$\begin{aligned} Exp(U(A)) &= \Pr(p)U(Ap) + \Pr(\neg p)U(A\neg p) \\ &= \Pr(p)(1 - \Pr(p)) + \Pr(\neg p)U(A\neg p) \\ &= \Pr(p)(1 - \Pr(p)) + (1 - \Pr(p))(-\Pr(p)) \\ &= \Pr(p)(1 - \Pr(p)) - (1 - \Pr(p))(\Pr(p)) \\ &= 0 \end{aligned}$$

So if you pay \$Pr(p) for the bet, your expected return is exactly 0. Obviously if you pay more, you're worse off, and if you pay less, you're better off. \$Pr(p) is the break even point, so that's the fair price for the bet.

And that's how we measure degrees of belief. We look at the agent's 'fair price' for a bet that returns \$1 if p. (Alternatively, we look at the maximum they'll pay for such a bet.) And that's their degree of belief that p. If we're taking probabilities to be degrees of belief, if we are (as it is sometimes put) interpreting probability subjectively, then that's the probability of p.

This might look suspiciously circular. The expected utility rule was meant to give us guidance as to how we should make decisions. But the rule needed a probability as an input. And now we're taking that probability to not only be a subjective state of the agent, but a subjective state that is revealed in virtue of the agent's own decisions. Something seems odd here.

Perhaps we can make it look even odder. Let p be some proposition that might be true and might be false, and assume that the agent's choice is to take or decline a bet on p that has some chance of winning and some chance of losing. Then if the agent takes the bet, that's a sign that their degree of belief in p was higher than the odds of the bet on p, so therefore they are increasing their expected utility by taking the bet, so they are doing the right thing. On the other hand, if they decline the bet, that's a sign that their degree of belief in p was lower than the odds of the bet on p, so therefore they are increasing their expected utility by taking the bet, so they are doing the right thing. So either way, they do the right thing. But a rule that says they did the right thing whatever they do isn't much of a rule.

There are two important responses to this, which are related to one another. The first is that although the rule does (more or less) put no restrictions at all on what you do when faced with a single choice, it can put quite firm constraints on your sets of choices when you have to make multiple decisions. The second is that the rule should be thought of as a **procedural** rather than **substantive** rule of rationality. We'll look at these more closely.

If we take probabilities to be subjective probabilities, i.e. degrees of belief, then the maximise expected utility rule turns out to be something like a consistency constraint. Compare it to a rule like *Have Consistent Beliefs*. As long as we're talking about logically contingent matters, this doesn't put any constraint at all on what you do when faced with a single question of whether to believe p or $\neg p$. But it does put constraints on what further beliefs you can have once you believe p. For instance, you can't now believe $\neg p$.

The maximise expected utility rule is like this. Indeed we already saw this in the Allais paradox. The rule, far from being empty, rules out the pair of choices that many people intuitively think is best. So if the objection is that the rule has no teeth, that objection can't hold up.

We can see this too in simpler cases. Let's say I offer the agent a ticket that pays \$1 if p, and she pays 60c for it. So her degree of belief in p must be at least 0.6. Then I offer her a ticket that pays \$1 if $\neg p$, and she pays 60c for it too. So her degree of belief in $\neg p$ must be at least 0.6. But, and here's the constraint, we think degrees of belief have to be probabilities. And if $\Pr(p) > 0.6$, then $\Pr(\neg p) < 0.4$. So if $\Pr(\neg p) > 0.6$, we have an inconsistency. That's bad, and it's the kind of badness it is the job of the theory to rule out.

One way to think about the expected utility rule is to compare it to norms of **means-end rationality**. At times when we're thinking about what someone should do, we really focus on what the best means is to their preferred end. So we might say *If you want to go to Harlem, you should take the A train*, without it even being a relevant question whether they should, in the circumstances, want to go to Harlem.

The point being made here is quite striking when we consider people with manifestly crazy beliefs. If we're just focussing on means to an end, then we might look at someone who, say, wants to crawl from the southern tip of Broadway to its northern tip. And we'll say "You should get some kneepads so you don't scrape your knees, and you should take lots of water, and you should catch the 1 train down to near to where Broadway starts, etc." But if we're not just offering procedural advice, but are taking a more substantive look at their position, we'll say "You should come up with a better idea about what to do, because that's an absolutely crazy thing to want."

As we'll see, the combination of the maximise expected utility rule with the use of degrees of belief as probabilities leads to a similar set of judgments. On the one hand, it is a very good guide to procedural questions. But it leaves some substantive questions worryingly unanswered. That has prompted some people to look for a more substantive characterisation of probability.

4.3 Credences and Norms

The previous section raised two large questions.

- Do we get the *right* procedural/consistency constraints from the expected utility rule? In particular (a) should credences be probabilities, and (b) should we make complex decisions by the expected utility rule? We'll look a bit in what follows at each of these questions.
- Is a purely procedural constraint all we're looking for in a decision theory?

And intuitively the answer to the second question is **No**. Let's consider a particular case. Alex is very confident that the Kansas City Royals will win baseball's World Series next year. In fact, Alex's credence in this is 0.9, very close to 1. Unfortunately, there is little reason for this confidence. Kansas City has been one of the worst teams in baseball for many years, the players they have next year will be largely the same as the players they had when doing poorly this year, and many other teams have players who have performed much much better. Even if Kansas City were a good team, there are 30 teams in baseball, and relatively random events play a big role in baseball, making it unwise to be too confident that any one team will win.

Now, Alex is offered a bet that leads to a $1 win if Kansas City win the World Series, and a $1 loss if they do not. The expected return of that bet, given Alex's credences, is $+80c$. So should Alex make the bet?

Intuitively, Alex should not. It's true that given Alex's credences, the bet is a good one. But it's also true that Alex has crazy credences. Given more sensible credences, the bet has a negative expected return. So Alex should not make the bet.

It's worth stepping away from probabilities, expected values and the like to think about this in a simpler context. Imagine a person has some crazy beliefs about what is an effective way to get some good end. And assume they, quite properly, want that good end. In fact, however, acting on their crazy beliefs will be counterproductive; it will just make things worse for everyone. And their evidence supports this. Should they act on their beliefs? Intuitively not. To be sure, if they didn't act on their beliefs, there would be some inconsistency between their beliefs and their actions. But inconsistency isn't the worst thing in the world. They should, instead, have different beliefs.

Similarly Alex should have different credences in the case in question. The question, what should Alex do given these credences, seems less interesting than the question, what should Alex do? And that's what we'll look at.

4.4 Evidential Probability

We get a better sense of what an agent should do if we look not to what credences they have, but to what credences they *should* have. Let's try to formalise this as the credences they would have if they were perfectly rational.

Remember credences are still being measured by betting behaviour, but now it is betting behaviour under the assumption of perfect rationality. So the probability of p is the highest price the agent would pay for a bet that pays \$1 if p, if they were perfectly rational. The thing that should be done then is the thing that has the highest expected utility, relative to this probability function. In the simple case where the choice is between taking and declining a bet, this becomes a relatively boring theory - you should take the bet if you would take the bet if you were perfectly rational. In the case of more complicated decisions, it becomes a much more substantive theory. (We'll see examples of this in later weeks.)

But actually we've said enough to give us two philosophical puzzles.

The first concerns whether there determinately is a thing that you would do if you were perfectly rational. Consider a case where you have quite a bit of evidence for and against p. Different rational people will evaluate the evidence in different ways. Some people will evaluate p as being more likely than not, and so take a bet at 50/50 odds on p. Others will consider the evidence against p to be stronger, and hence decline a bet at 50/50 odds. It seems possible that both sides in such a dispute could be perfectly rational.

The danger here is that if we define rational credences as the credences a perfectly rational person would have, we might not have a precise definition. There may be many different credences that a perfectly rational person would have. That's bad news for a purported definition of rational credence.

The other concerns cases where p is about your own rationality. Let's say p is the proposition that you are perfectly rational. Then if you were perfectly rational, your credence in this would probably be quite high. But that's not the rational credence for you to have right now in p. You should be highly confident that you, like every other human being on the planet, are susceptible to all kinds of failures of rationality. So it seems like a mistake in general to set your credences to what they would be were you perfectly rational.

What seems better in general is to proportion your credences to the evidence. The rational credences are the ones that best reflect the evidence you have in favour of various propositions. The idea here to to generate what's usually called an **evidential probability**. The probability of each proposition is a measure of how strongly it is supported by the evidence.

That's different from what a rational person would believe in two respects. For one thing, there is a fact about how strongly the evidence supports p, even if different people might disagree about just how strongly that is. For another thing, it isn't true that the evidence supports that you are perfectly rational, even though you would believe that if you were perfectly rational. So the two objections we just mentioned are not an issue here.

From now on then, when we talk about probability in the context of expected utility, we'll talk about evidential probabilities. There's an issue about whether we can numerically measure strengths of evidence. That is, there's an issue about whether strengths of evidence are the right kind of thing to be put on a numerical scale. Even if they are, there's a tricky issue about how we can even guess what they are. I'm going to cheat a little here. Despite the arguments above that evidential probabilities can't be *identified* with betting odds of perfectly rational agents, I'm going to assume that, unless we have reason to the contrary, those betting odds will be our first approximation. So when we have to guess what the evidential probability of p is, we'll start with what odds a perfectly rational agent (with your evidence) would look for before betting on p.

4.5 Objective Chances

There is another kind of probability that theorists are often interested in, one that plays a particularly important role in modern physics. Classical physics was, or at least was thought to be, deterministic. Once the setup of the universe at a time t was set, the laws of nature determined what would happen after t. Modern physics is not deterministic. The laws don't determine, say, how long it will take for an unstable particle to decay. Rather, all the laws say is that the particle has such-and-such a chance of decaying in a certain time period. You might have heard references to the half-life of different radioactive particles; this is the time in which the particle has a $\frac{1}{2}$ probabiilty of decaying.

What are these probabilities that the scientists are talking about? Let's call them 'chances' to give them a name. So the question is, what is the status of chance?. We know chances aren't evidential probabilities. We know this for three reasons.

One is that it is a tricky empirical question whether any event has any chance other than 0 or 1. It is now something of a scientific consensus that some events are indeed chancy. But this relies on some careful scientific investigation. It isn't something we can tell from our armchairs. But we can tell from just thinking about decisions under uncertainty that for at least some outcomes the evidential probability of that outcome is between 0 and 1.

Another is that, as chances are often conceived, events taking place in the past do not, right now, have chances other than 0 or 1. There might have been, at a point in the past, some intermediate chance of a particle decaying. But if we're now asking about whether a particle did decay or not in the last hour, then either it did decay, and its chance is 0, or it did not decay, and its chance is 1. (I should note that not everyone thinks about chances in quite this way, but it is a common way to think about them.) There are many events that took place in the past, however, whose evidential probability is between 0 and 1. For instance, if we're trying to meet up a friend, and hence trying to figure out where the friend might have gone to, we'll think about, and assign evidential probabilities to, various paths the friend might have taken in the past. These thoughts won't be thoughts about chances in the physicists' sense; they'll be about evidential probabilities.

Finally, chances are objective. The evidential probability that p is true might be different for me than for you. For instance, the evidence she has might make it quite likely for the juror that the suspect is guilty, even if he is not. But the evidence the suspect has makes it extremely likely that he is innocent. Evidential probabilities differ between different people. Chances do not. Someone might not know what the chance of a particular outcome is, but what they are ignorant of is a matter of objective fact.

The upshot seems to be that chances are quite different things from evidential probabilities, and the best thing to do is simply to take them to be distinct basic concepts.

4.6 The Principal Principle and Direct Inference

Although chances and evidential probabilities are distinct, it seems they stand in some close relation. If a trustworthy physicist tells you that a particle has an 0.8 chance of decaying in the next hour, then it seems your credences should be brought into line with what the physicists say. This idea has been dubbed the Principal Principle, because it is the main principle linking chances and credences. If we use Pr for evidential probabilities, and Ch for objective chances in the physicists' sense, then the idea behind the principle is this.

Principal Principle $\Pr(p|Ch(p)=x)=x$

That is, the probability of p, conditional on the chance of p being x, is x.

The Principal Principle may need to be qualified. If your evidence also includes that p, then even if the chance of p is 0.8, perhaps your credence in p should be 1. After all, p is literally evident to you. But perhaps it is impossible for p to be part of your evidence while its chance is less than 1. The examples given in the literature of how this could come about are literally spectacular. Perhaps God tells you that p is true. Or perhaps a fortune teller with a crystal ball sees that it is true. Or something equally bizarre happens. Any suggested exceptions to the principle have been really outlandish. So whether the principle is true for all possible people in all possible worlds, it seems to hold for us around here.

Chances, as the physicists think of them, are not frequencies. It might be possible to compute the theoretical chance of a rare kind of particle not decaying over the course of an hour, even though the particle is so rare, and so unstable, that no such particle has ever survived an hour. In that case the frequency of survival (i.e. the proportion of all such particles that do actually survive an hour) is 0, but physical theory might tell us that the chance is greater than 0. Nevertheless chances are like frequencies in some respects.

One such respect is that chances are objective. Just as the chance of a particle decay is an objective fact, one that we might or might not be aware of, the frequency of particle decay is also an objective fact that we might or might not be aware of. Neither of these facts are in any way relative to the evidence of a particular agent, the way that evidential probabilities are.

And just like chances, frequencies might seem to put a constraint on credences. Consider a case where the only thing you know about a is that it is G. And you know that the frequency of F-hood among Gs is x. For instance, let a be a person you've never met, G be the property of being a 74 year old male smoker, and F the property of surviving 10 more years. Then you might imagine knowing the survival statistics, but knowing nothing else about the person. In that case, it's very tempting to think the probability that a is F is x. In our example, we'd be identifying the probability of this person surviving with the frequency of survival among people of the same type.

This inference from frequencies to probabilities is sometimes called "Direct Inference". It is, at least on the surface, a lot like the Principal Principle. But it is a fair bit more contentious. It is really rather rare that all we know about an individual can be summed up in one statistic like this. Even if the direct inference can be philosophically justified (and I'm a little unsure that it can be) it will rarely be applicable. So it is less important than the Principal Principle.

We'll often invoke the Principal Principle tacitly in setting up problems. That is, when I want to set up a problem where the probabilities of the various outcomes are given, I'll often use objective chances to fix the probabilities of various states. We'll use the direct inference more sparingly, because it isn't as clearly useful.

4.7 Money and Utility

Let's change tack a little, and think a bit more about utility rather than probability. In simple puzzles involving money, it is easy to think of the dollar amounts involved as being proxy for the utility of each outcome. In a lot of cases, that's a very misleading way of thinking about things though. In general, a certain amount of money will be less useful to you if you have more money. So \$1000 will be more useful to a person who earns \$20,000 per year than a person who earns \$100,000 per year. And \$1,000,000 will be more useful to either of them than it will be to, say, Bill Gates.

This matters for decision making. It matters because it implies that in an important sense, $\$2x$ is generally not twice as valuable to you as $\$x$. That's because $\$2x$ is like getting $\$x$, and then getting $\$x$ again. (A lot like it really!) And when we're thinking about the utility of the second $\$x$, we have to think about its utility not to you, but to the person you'll be once you've already got the first $\$x$. And that person might not value the second $\$x$ that much.

To put this in perspective, consider having a choice between \$1,000,000 for certain, and a 50% chance at \$2,000,000. Almost everyone would take the sure million. And that would be rational, because it has a higher utility. It's a tricky question to think about just what is the smallest x for which you'd prefer a 50% chance at $\$x$ to \$1,000,000. It might be many many times more than a million.

The way economists' put this is that money (like most goods) has a *declining marginal utility*. The marginal utility of a good is, roughly, the utility of an extra unit of the good. For a good like money that comes in (more or less) continuous quantities, the marginal utility is the slope of the utility graph, as below.

You should read the x-axis there are measuring possible incomes in thousands of dollars per year, and the y-axis as measuring utility. The curve there is $y = x^{\frac{1}{2}}$. That isn't necessarily a plausible account of how much utility each income might give you, but it's close enough for our purposes. Note that although more income gives you more utility, the amount of extra utility you get from each extra bit of income goes down as you get more income. More precisely, the slope of the income-utility graph keeps getting shallower and shallower as your

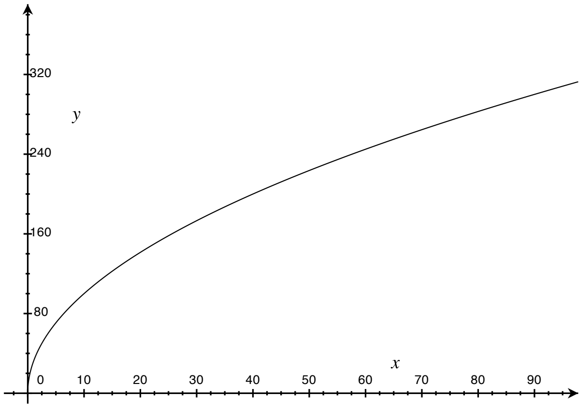


Figure 4.1: Declining Marginal Utility of Money

income/utility rises. (More precisely yet, a little calculus shows that the slope of the graph at any point is $\frac{1}{2y}$, which is obviously always positive, but gets less and less as your income/utility gets higher and higher.)

The fact that there is a declining marginal utility of money explains certain features of economic life. We'll look at models of two simple economic decisions, buying insurance and diversifying an investment portfolio. We'll then use what we said about diversified investments to explain some features of the actual insurance markets that we find.

4.8 Insurance

Imagine the utility an agent gets from an income of x dollars is $x^{\frac{1}{2}}$. And imagine that right now their income is \$90,000. But there is a 5% chance that something catastrophic will happen, and their income will be just \$14,400. So their expected income is $0.95 \times 90,000 + 0.05 \times 14,400 = 86220$. But their expected utility is just $0.95 \times 300 + 0.05 \times 120 = 291$, or the utility they would have with an income of \$84,861.

Now imagine this person is offered insurance against the catastrophic scenario. They can pay, say, \$4,736, and the insurance company will restore the \$75,600 that they will lose if the catastrophic event takes place. Their income is now sure to be \$85,264 (after the insurance is taken out), so they have a utility of 292. That's higher than what their utility was, so this is a good deal for them.

But note that it might also be a good deal for the insurance company. They receive in premiums \$4,736. And they have a 5% chance of paying out \$75,600. So the expected outlay, in dollars, for them, is \$3,780. So they turn an expected profit on the deal. If they repeat this deal often enough, the probability that they will make a profit goes very close to 1.

The point of the example is that people are trying to maximise expected utility, while insurance companies are trying to maximise expected profits. Since there are cases where lowering your expected income can raise your expected utility, there is a chance for a win-win trade. And this possibility, that expected income can go down while expected utility can go up, is explained in virtue of the fact that there is a declining marginal utility of money.

4.9 Diversification

Imagine that an agent has a starting wealth of 1, and the utility the agent gets from wealth x is $x^{\frac{1}{2}}$. (We won't specify 2 what, but take this to be some kind of substantial unit.) The agent has an opportunity to make an investment that has a 50% chance of success and a 50% chance of failure. If the agent invests y in the scheme, the returns r will be

$$r = \begin{cases} 4y, & \text{if success,} \\ 0, & \text{if failure.} \end{cases}$$

The expected profit, in money, is y. That's because there is a 50% chance of the profit being $3y$, and a 50% chance of it being $-y$. But in utility, the expected return of investing 1 unit is 0. The agent has a 50% chance of ending with a wealth of 4, i.e. a utility of 2, and a 50% chance of ending with a wealth of 0, i.e. a utility of 0.

So making the investment doesn't seem like a good idea. But now imagine that the agent could, instead of putting all their money into this one venture, split the investment between two ventures that (a) have the same probability of returns as this one, and (b) their success of failure is probabilistically independent. So the agent invests $\frac{1}{2}$ in each deal. The agent's return will be

$$r = \begin{cases} 4, & \text{if both succeed,} \\ 2, & \text{if one succeeds and the other fails,} \\ 0, & \text{if both fail.} \end{cases}$$

The probability that both will succeed is $\frac{1}{4}$. The probability that one will succeed and the other fail is $\frac{1}{2}$. (Exercise: why is this number greater?) The probability that both will fail is $\frac{1}{4}$. So the agent's expected profit, in wealth, is 1. That is, it is $4 \times \frac{1}{4} + 2 \times \frac{1}{2} + 0 \times \frac{1}{4}$, i.e. 2, minus the 1 that is invested, so it is 2 minus 1, i.e. 1. So it's the same as before. Indeed, the expected profit on each investment is $\frac{1}{2}$.

And the expected profits on a pair of investments is just the sum of the expected profits on each of the investments.

But the expected utility of the 'portfolio' of two investments is considerably better than other portfolios with the same expected profit. One such portfolio is investing all of the starting wealth in one 50/50 scheme. The expected utility of the portfolio is $4^{\frac{1}{2}} \times \frac{1}{4} + 2^{\frac{1}{2}} \times \frac{1}{2} + 0 \times \frac{1}{4}$, which is about 1.21. So it's a much more valuable portfolio to the agent than the portfolio which had just a single investment. Indeed, the diversified investment is worth making, while the single investment was not worth making.

This is the general reason why it is good to have a diversified portfolio of investments. It isn't because the expected profits, measured in dollars, are higher this way. Indeed, diversification couldn't possibly produce a higher expected profit. That's because the expected profit of a portfolio is just the sum of the expected profits of each investment in the portfolio. What diversification can do is increase the expected utility of that return. Very roughly, the way it does this is by decreasing the probability of the worst case scenarios, and of the best case scenarios. Because the worst case scenario is more relevant to the expected utility calculation than the best case scenario, because in general it will be further from the median outcome, the effect is to increase the expected utility overall.

One way of seeing how important diversification is is to consider what happens if the agent again makes two investments like this, but the two investments are probabilistically linked. So if one investment succeeds, the other has an 80% chance of success. Now the probability that both will succeed is 0.4, the probability that both will fail is 0.4, and the probability that one will succeed and the other fail is 0.2. The expected profit of the investments is still 1. (Each investment still has an expected profit of $\frac{1}{2}$, and expected profits are additive.) But the expected utility of the portfolio is just $4^{\frac{1}{2}} \times 0.4 + 2^{\frac{1}{2}} \times 0.2 + 0 \times 0.4$, which is about 1.08. The return on investment, in utility terms, has dropped by more than half.

The lesson is that for agents with declining marginal utilities for money, a diversified portfolio of investments can be more valuable to them than any member of the portfolio on its own could be. But this fact turns on the investments being probabilistically separated from one another.

4.10 Selling Insurance

In the toy example about insurance, we assumed that the marginal utility of money for the insurance company was flat. That isn't really true. The insurance company is owned by people, and the utility of return to those people is diminishing as the returns get higher. There is also the complication that the insurance company faces very different kinds of returns when it is above and below the solvency line.

Nevertheless, the assumption that the marginal utility of money is constant for the insurance company is constant is a useful fiction. And the reason that it is a useful fiction is that if the insurance company is well enough run, then the assumption is close to being true. By 'well enough run', I simply mean that their insurance portfolio is highly diversified.

We won't even try to prove this here, but there are various results in probability theory that suggest that as long as there are a lot of different, and probabilistically independent, investments in a portfolio, then with a very high probability, the actual returns will be close to the expected returns. In particular, if the expected returns are positive, and the portfolio is large and diverse enough, then with a very high probability the actual returns will be positive. So, at least in optimal cases, it isn't a terrible simplification to treat the insurance company as if it was sure that it would actually get its expected profits. And if that's the case, the changing marginal utility of money is simply indifferent.

The mathematical results that are relevant here are what are sometimes called the "Law of Large Numbers". The law says that if you sample independent and identically distributed random variables repeatedly, then for any positive number e, the probability that the average output is within e of the expected output goes to 1 as the number of samples goes to infinity. The approach can be quite quick in some cases. The following table lists the probability that the number of heads on n flips of a random coin will be (strictly) between $0.4n$ and $0.6n$ for various values of n.

Number of flips	Probabiilty of between $0.4n$ and $0.6n$ heads
1	0
10	0.246
20	0.497
50	0.797
100	0.943
200	0.994
500	> 0.99

This depends crucially on independence. If the coin flips were all perfectly dependent, then the probabilities would not converge at all.

Note we've made two large assumptions about insurance companies. One is that the insurance company is large, the other is that it is diversified. Arguably both of these assumptions are true of most real-world insurance companies. There tend to be very few insurance companies in most economies. More importantly, those companies tend to be fairly diversified. You can see this in a couple of different features of modern insurance companies.

One is that they work across multiple sectors. Most car insurance companies will also offer home insurance. Compare this to other industries. It isn't common for car sales agents to also be house sales agents. And it isn't common for car builders to also be house builders. The insurance industry tends to be special here. And that's because it's very attractive for the insurance companies to have somewhat independent business wings, such as car insurance and house insurance.

Another is that the products that are offered tend to be insurance on events that are somewhat probabilistically independent. If I get in a car accident, this barely makes a difference to the probability that you'll be in a car accident. So offering car insurance is an attractive line of business. Other areas of life are a little trickier to insure. If I lose my home to a hurricane, that does increase, perhaps substantially, the probability of you losing your house to a hurricane. That's because the probability of their being a hurricane, conditional on my losing my house to a hurricane, is 1. And conditional on their being a hurricane, the probability of you losing your house to a hurricane rises substantially. So offering hurricane insurance isn't as attractive a line of business as car insurance. Finally, if I lose my home to an invading army, the probability that the same will happen to you is very high indeed. In part for that reason, very few companies ever offer 'invasion insurance'.

It is very hard to say with certainty whether this is true, but it seems that a large part of the financial crisis that has been going for the last few years is related to a similar problem. A lot of the financial institutions that failed were selling, either explicitly or effectively, mortgage insurance. That is, they were insuring various banks against the possibility of default. One problem with this is that mortgage defaults are not probabilistically independent. If I default on my mortgage, that could be because I lost my job, or it could be because my house price collapsed and I have no interest in sustaining my mortgage. Either way, the probability that you will also default goes up. (It goes up dramatically if I defaulted for the second reason.) What may have been sensible insurance policies to write on their own turned into massive losses because the insurers underestimated the probability of having to pay out on many policies all at once.

Chapter 5

Utility

5.1 Utility and Welfare

So far we've frequently talked about the utility of various outcomes. What we haven't said a lot about is just what it is that we're measuring when we measure the utility of an outcomes. The intuitive idea is that utility is a measure of welfare - having outcomes with higher utility if a matter of having a higher level of welfare. But this doesn't necessarily move the idea forward, because we'd like to know a bit more about what it is to have more welfare. There are a number of ways we can frame the same question. We can talk about 'well-being' instead of welfare, or we can talk about having a good life instead, or having a life that goes well. But the underlying philosophical question, what makes it the case that a life has these features, remains more or less the same.

There are three primary kinds of theories of welfare in contemporary philosophy. These are

- Experience Based theories
- Objective List theories
- Preference Based theories

In decision theory, and indeed in economics, people usually focus on preference based theories. Indeed, the term 'utility' is sometimes used in such way that A has more utility than B just means that the agent prefers A to B. Indeed, I've sometimes earlier moved back and forth previously between saying A has higher utility and saying A is preferred. And the focus here (and in the next set of notes) will be on why people have moved to preference based accounts, and technical challenges within those accounts. But we'll start with the non-preference based accounts of welfare.

5.2 Experiences and Welfare

One tradition, tracing back at least to Jeremy Bentham, is to identify welfare with having good experiences. A person's welfare is high if they have lots of pleasures, and few pains. More generally, a person's welfare is high if they have good experiences.

Of course it is possible that a person might be increasing their welfare by having bad experiences at any one time. They might be at work earning the money they need to finance activities that lead to good experiences later, or they might just be looking for money to stave off bad experiences (starvation, lack of shelter) later. Or perhaps the bad experiences, such as in strenuous exercise, are needed in order to be capable of later doing the things, e.g. engaging in sporting activities, that produce good experiences. Either way, the point has to be that a person's welfare is not simply measured by what their experiences are like right now, but by what their experiences have been, are, and will be over the course of their lives.

There is one well known objection to any such account - what Robert Nozick called the "experience machine". Imagine that a person is, in their sleep, kidnapped and wired up to a machine that produces in their brain the experiences as of a fairly good life. The person still seems to be having good days filled with enjoyable experiences. And they aren't merely raw pleasurable sensations - the person is having experiences as of having rich fulfilling relationships with the friends and family they have known and loved for years. But in fact the person is not in any contact with those people, and for all the friends and family know, the person was kidnapped and killed. This continues for decades, until the person has a peaceful death at an advanced age.

Has this person had a good life or a bad life? Many people think intuitively that they have had a bad life. Their entire world has been based on an illusion. They haven't really had fulfilling relationships, travelled to exciting places, and so on. Instead they have been systematically deceived about the world. But on an experience based view of welfare, they have had all of the goods you could want in life. Their experiences are just the experiences that a person having a good life would have. So the experience based theorist is forced to say that they have had a good life, and this seems mistaken.

Many philosophers find this a compelling objection to the experience based view of welfare. But many people are not persuaded. So it's worth thinking a little through some other puzzles for purely experience based views of welfare.

It's easy enough to think about paradigmatic pains, or bad experiences. It isn't too hard to come up with paradigmatic good experiences, though perhaps there would be more disagreement about what experiences are paradigms of the good than are paradigms of the bad. But many experiences are less easy to classify. Even simple experiences like tickles might be good experiences for some, and bad experiences for others.

When we get to more complicated experiences, things are even more awkward for the experience based theorist. Some people like listening to heavily distorted music, or watching horror movies, or drinking pineapple schnapps. Other people, indeed most people, do not enjoy these things. The experience theory has a couple of choices here. Either we can say that one group is wrong, and these things either do, or do not, raise one's welfare. But this seems implausible for all experiences. Perhaps at the fringes there are experiences people seek that nevertheless decrease their welfare, but it seems strange to argue that the same experiences are good for everyone.

The other option is to say that there are really two experiences going on when you, say, listen to a kind of music that some, but not all, people like. There is a 'first-order' experience of hearing the music. And there is a 'second-order' experience, an experience of enjoying the experience of hearing the music. Perhaps this is right in some cases. (Perhaps for horror movies, fans both feel horrified and have a pleasant reaction to being horrified, at least some of the time.) But it seems wrong in general. If there is a food that I like and you dislike, that won't usually be because I'll have a positive second-order experience, and you won't have such a thing. Intuitively, the experience of, say, drinking a good beer, isn't like that, because it just isn't that complicated. Rather, I just have a certain kind of experience, and I like it, and you, perhaps, do not.

A similar problem arises when considering the choices people make about how to distribute pleasures over their lifetime. Some people are prepared to undergo quite unpleasant experiences, e.g. working in painful conditions, in exchange for pleasant experiences later (e.g. early retirement, higher pay, shorter hours). Other people are not. Perhaps in some cases people are making a bad choice, and their welfare would be higher if they made different trade-offs. But this doesn't seem to be universally true - it just isn't clear that there's such a thing as the universally correct answer to how to trade off current unpleasantness for future pleasantness.

Note that this intertemporal trade-off question actually conceals two distinct questions we have to answer. One is how much we want to 'discount' the future. Economists think, with some empirical support, that people mentally discount future goods. People value a dollar now more than they value a dollar ten years hence, or even an inflation adjusted dollar ten years hence. The same is true for experiences: people value good experiences now more than good experiences in the future. But it isn't clear how much discount, if any, is consistent with maximising welfare. The other question is how much we value high 'peaks' of experience versus avoiding low 'troughs'. Some people are prepared to put up with the bad to get the good, others are not. And the worry for the experience based theorist is that neither need be making a mistake. Perhaps what is best for a person isn't just a function of their experiences over time, but on how much they value the kind of experiences that they get.

So we've ended up with three major kinds of objections to experience based accounts of welfare.

- The experience machine does not increase our welfare
- Different people get welfare from different experiences
- Different people get different amounts of welfare from the same sequences of experiences over time, even if they agree about the welfare of each of the moment-to-moment experiences.

These seem like enough reasons to move to other theories of welfare.

5.3 Objective List Theories

One response to these problems with experience based accounts is to move to a theory based around desire satisfaction. Since that's the theory that's most commonly used in decision theory, we'll look at it last. Before that, we'll look briefly at so called *objective list* theories of welfare. These theories hold that there isn't necessarily any one thing that makes your life better. Welfare isn't all about good experiences, or about having preferences that are satisfied. Rather, there are many ways in which your welfare can be improved. The list of things that make your life better may include:

- Knowledge
- Engaging in rational activity
- Good health, adequate shelter, and more generally good physical well-being
- Being in loving relationships, and in sustained friendships
- Being virtuous
- Experiencing beauty
- Desiring the things that make life better, i.e. the things on this list

Some objective list theorists hold that the things that should go on the list do have something in common, but this isn't an essential part of the theory.

The main attraction of the objective list approach is negative. We've already seen some of the problems with experience based theories of welfare. We'll see later some of the problems with desire based theories. A natural response to this is to think that welfare is heteroegenous, and that no simple theory of welfare can capture all that makes human lives go well. That's the response of the objective list theorist.

The first thing to note about these theories is that the lists in question always seem open to considerable debate. If there was a clearer principle about what's going on the lists and what is not, this would not be such a big deal. But in the absence of a clear (or easy to apply) principle, there is a sense of arbitrariness about the process.

Indeed, the lists that are produced by Western academics seem notably aligned with the desires and values of Western academics. It's notable that the lists produced tend to give very little role to the family, to religion, to community and to tradition. Of course all these things can come in indirectly. If being in loving relationships is a good, and families promote loving relationships, then families are an indirect good. And the same thing can be said religion, and community, and traditional practices. But still, many people might hold those things to be valuable in their own right, not just because of the goods that they produce. Or they might hold some things on the canonical lists, such as education and knowledge to be instrumental goods, rather than making them primary goods as philosophers often do.

This can't be an objection to objective list theories of welfare as such. Nothing in the theory rules out extending the list to include families, or traditions, in the mix, for instance. (Indeed, these kinds of goods are included in some versions of the theory.) But it is perhaps revealing that the lists hew so closely to the Western academic's idea of the good life. (Indeed the list I've got here is more universal than several proposed lists, since I've included health and shelter, which is left off some.) It might well be thought that there isn't one list of goods that make life good for any person in any community at any time. There might well be a list of what makes for a good life in a community like ours, and maybe even lists like the one above capture it, but claims to universality should be treated sceptically.

A more complicated question is how to generate comparative welfare judgments from the list. Utilities are meant to be represented numerically, so we need to be able to say which of two outcomes is better, or that the outcomes are exactly as good as one another. (Perhaps we need something more, some way of saying how much better one life is than another. But we'll set that question aside for now.) We already saw one hard aspect of this question above - how do we turn facts about the welfare of a person at different times of their life into an overall welfare judgment? That question is just as hard for the objective list theorist as for the experience theorist. (And again, part of why it is so hard is that it is far from clear that there is a unique correct answer.)

But the objective list theorist has a challenge that the experience theorist does not have: how do we weigh up the various goods involved? Let's think about a very simple list - say the only things on the list are friendship and beauty. Now in some cases, saying which of two outcomes is better will be easy. If outcome

A will produce improve your friendship, and let you experience beautiful things, more than outcome *B* will, then *A* is better than *B*. But not all choices are like that. What if you are faced with a choice between seeing a beautiful art exhibit, that is closing today, or keeping a promise to meet your friend for lunch? Which choice will maximise your welfare? The art gallery will do better from a beauty standpoint, while the lunch will do better from a friendship standpoint. We need to know something more to know how this tradeoff will be made.

There are actually three related objections here. One is that the theory is incomplete unless there is some way to weigh up the various things on the list, and the list itself does not produce the means to do the weighting. A second is that it isn't obvious that there is a unique way to weigh up the things on the list. Perhaps one person is made better off by focussing on friendship and the expense of beauty, and for another person it goes the other way. So perhaps there is no natural weighing consistent with the spirit behind the objective list theories that works in all contexts. Finally, it isn't obvious that there is a fact of the matter in many cases, leaving us with many choices where there is no fact of the matter about which will produce more utility. But that will be a problem for creating a numerical measure of value that can be plugged into expected utility calculations.

Let's sum up. There are really two core worries about objective list theories. These are:

- Different things are good for different people
- There's no natural way to produce a utility measure out of the goodness of each 'component' of welfare

Next we'll look at desire based theories of utility, which are the standard in decision theory and in economics.

5.4 Preference Based Theories

So far we've looked at two big theories of the nature of preferences. Both of them have thought that in some sense people don't get a say in what's good for them. There is an impersonal fact about what is best for a person, and that is good for you whether you like it or not. The experience theory says that it is the having of good experiences, and the objective list theory says that it includes a larger number of features. Preference-based, or 'subjective' theories of welfare start with the idea that what's good for different people might be radically different. It also takes the idea that people often are the best judge of what's best for them very seriously.

What we end up with is the theory that *A* is better for an agent than *B* if and only if the agent prefers *A* to *B*. We'll look at some complications to this, but for now we'll work with the simple picture that welfare is a matter of preference satisfaction. This theory has a number of advantages over the more objective theories.

First, it easily deals with the idea that different things might be good for different people. That's accommodated by the simple fact that people have very different desires, so different things increase their welfare.

Second, it also deals easily with the issues about comparing bundles of goods, either bundles of different goods, or bundles of goods at different times. An agent need not only have preferences about whether they, for instance, prefer time with their family to material possessions. They also have more fine-grained preferences about various trade offs between different goods, and trade offs about sequences of goods across time. So if one person has a strong preference for getting goods now, and another person is prepared to wait for greater goods later, the theory can accommodate that difference. Or if one person is prepared to put up with unpleasant events in order to have greater goods at other times, the theory can accommodate that, as well as the person who prefers a more steady life. If they are both doing what they want, then even though they are doing different things, they are both maximising their welfare.

But there are several serious problems concerning this approach to welfare. We'll start with the intuitive idea that people sometimes don't know what is good for them.

We probably all can think about things in everyday life where we, or a friend of ours, has done things that quite clearly are not in their own best interests. In many such cases, it won't be that the person is doing what they don't want to do. Indeed, part of the reason that people acting against their own best interests is such a problem is that the actions in question are ones they very much want to perform. Or so we might think antecedently. If a person's interests are just measured by their desires, then it is impossible to want what's bad for you. That seems very odd.

It is particularly odd when you think about the effect of advertising and other forms of persuasion. The point of advertising is to change your preferences, and presumably it works frequently enough to be worth spending a lot of money on. But it is hard to believe that the effect of advertising is to change how good for you various products are. Yet if your welfare is measured by how many of your desires are satisfied, then anything that changes your desires changes what is good for you.

Note that sometimes we even have internalised the fact that we desire the wrong things. Sometimes we desire something, while desiring that we don't desire it. So we can say things like "I wish I didn't want to smoke so much". In that case it seems that what would, on a strict subjective standpoint, have our best outcome be smoking and wanting not to smoke, since then both our ‘first-order’ desire to smoke and our ‘second-order’ desire not to want to smoke would be satisfied. But that sounds crazy.

Perhaps the best thing to do here would be to modify the subjective theory of welfare. Perhaps we could say that our welfare is maximised by the satisfaction of those desires we wish we had. Or perhaps we could say that it is maximised by the satisfaction of our ‘undefeated’ desires, i.e. desires that we don’t wish we didn’t have. There are various options here for keeping the spirit of a subjective approach to welfare, while allowing that people sometimes desire the bad.

5.5 Interpersonal Comparisons

I mentioned above that the subjective approach does better than the other approaches at converting the welfare someone gets from the different parts of their life into a coherent whole. That’s because agent’s don’t only have preferences over how the parts of their lives go, they also have preferences over different distributions of welfare over the different parts of their lives, and preferences over bundles of goods they may receive. The downside of this is that a kind of comparison that the objective theory might do well at, interpersonal comparisons, are very hard for the subjective theorist to make.

Intuitively there are cases where the welfare of a group is improved or decreased by a change in events. But this is hard, in general, to capture on a subjective theory of welfare. There is one kind of group comparison that we can make. If some individuals prefer A to B, and none prefer B to A, then A is said to be a Pareto-improvement over B. (The name comes from the Italian economist Wilfredo Pareto.) An outcome is Pareto-optimal if no outcome is a Pareto-improvement over it.

But Pareto-improvements, and even Pareto-inefficiency, are rare. If I’m trying to decide who to give \$1000 to, then pretty much whatever choice I make will be Pareto-optimal. Assume I give the money to x. Then any other choice will involve x not getting \$1000, and hence not preferring that outcome. So not everyone will prefer the alternative.

But intuitively, there are cases which are not Pareto-improvements which make a group better off. Consider again the fact that the marginal utility of money is declining. That suggests that if we took \$1,000,000 from Bill Gates, and gave \$10,000 each to 100 people on the borderline of losing their houses, then we’d have increased the net welfare. It might not be just to simply take money from Gates in this way, so many people will think it would be wrong to do even if it wouldn’t increase welfare. But it would be odd to say that this didn’t increase welfare. It might be odder still to say, as the subjective theory seems forced to say, that there's no way to tell whether it increased welfare, or perhaps that there is no fact of the matter about whether it increased net welfare, because welfare comparisons only make sense for something that has desires, e.g. an agent, not something that does not, e.g. a group.

There have been various attempts to get around this problem. Most of them start with the idea that we can put everyone's preferences on a scale with some fixed points. Perhaps for each person we can say that utility of 0 is where they have none of their desires satisfied, and utility of 1 is where they have all of their desires satisfied. The difficulty with this approach is that it suggests that one way to become very very well off is to have few desires. The easily satisfied do just as well as the super wealthy on such a model. So this doesn't look like a promising way forward.

Since we're only looking at decisions made by a single individual here, the difficulties that subjective theories of welfare have with interpersonal comparisons might not be the biggest concern in the world. But it is an issue that comes up whenever we try to apply subjective theories broadly.

5.6 Which Desires Count

There is another technical problem about using preferences as a foundation for utilities. Sometimes I'll choose A over B, not because A really will produce more welfare for me than B, but because I think that A will produce more utility. In particular, if A is a gamble, then I might take the gamble even though the actual result of A will be be worse, by anyone's lights, including my own, than B.

Now the subjectivist about welfare does want to use preferences over gambles in the theory. In particular, it is important for figuring out how much an agent prefers A to B to look at the agent's preferences over gambles. In particular, if the agent thinks that one gamble has a 50% chance of generating A, and a 50% chance of generating C, and the agent is indifferent between that gamble and B, then the utility of B is exactly half-way between A's utility and C's utility. That's a very useful thing to be able to say. But it doesn't help with the original problem - how much do we value actual outcomes, not gambles over outcomes.

What we want is a way of separating *instrumental* from *non-instrumental* desires. Most of our desires are, at least to some extent, instrumental. But that's a problem for using them in generating welfare functions. If I have an instrumental desire for A, that means I regard A as a gamble that will, under conditions I give a high probability of obtaining, lead to some result C that I want. What we really want to do is to specify these non-instrumental desires.

A tempting thing to say here is to look at our desires under conditions of full knowledge. If I know that the train and the car will take equally long to get to a destination I desire, and I still want to take the train, that's a sign that I have a genuine preference for catching the train. In normal circumstances, I might catch the train rather than take the car not because I have such a preference, but because I could be stuck in arbitrarily long traffic jams when driving, and I'd rather not take that risk.

But focussing on conditions of full knowledge won't get us quite the results that we want. For one thing, there are many things where full knowledge changes the relevant preferences. Right now I might like to watch a football game, even though this is something of a gamble. I'd rather do other things conditional on my team losing, but I'd rather watch conditional on them winning. But if I knew the result of the game, I wouldn't watch - it's a little boring to watch games where you know the result. The same goes of course for books, movies etc. And if I had full knowledge I wouldn't want to learn so much, but I do prefer learning to not learning.

A better option is to look at desires over fully specific options. A fully specific option is an option where, no matter how the further details are filled out, it doesn't change how much you'd prefer it. So if we were making choices over complete possible worlds, we'd be making choices over fully specific options. But even less detailed options might be fully specific in this sense. Whether it rains in an uninhabited planet on the other side of the universe on a given day doesn't affect how much I like the world, for instance.

The nice thing about fully specific options is that preferences for one rather than the other can't be just instrumental. In the fully specific options, all the possible consequences are played out, so preferences for one rather than another must be non-instrumental. The problem is that this is psychologically very unrealistic. We simply don't have that fine-grained a preference set. In some cases we have sufficient dispositions to say that we do prefer one fully specific option to another, even if we hadn't thought of them under those descriptions. But it isn't clear that this will always be the case.

To the extent that the subjective theory of welfare requires us to have preferences over options that are more complex than we have the capacity to consider, it is something of an idealisation. It isn't clear that this is necessarily a bad thing, but it is worth noting that the theory is in this sense a little unrealistic.

Chapter 6

Newcomb's Puzzle

This chapter will be primarily concerned wth the following puzzle, called Newcomb's Puzzle. (It is named for William Newcomb, who is responsible for its introduction to philosophical debate.) In front of you are two boxes, call them A and B. You call see that in box B there is $1000, but you cannot see what is in box A. You have a choice, but not perhaps the one you were expecting. Your first option is to take just box A, whose contents you do not know. Your other option is to take both box A and box B, with the extra $1000.

There is, as you may have guessed, a catch. A demon has predicted whether you will take just one box or take two boxes. The demon is very good at predicting these things – in the past she has made many similar predictions and been right every time. If the demon predicts that you will take both boxes, then she's put nothing in box A. If the demon predicts you will take just one box, she has put $1,000,000 in box A. So the table looks like this.

	Predicts 1 box	Predicts 2 boxes
Take 1 box	$1,000,000	$0
Take 2 boxes	$1,001,000	$1,000

There are interesting arguments for each of the two options here.

The argument for taking just one box is easy. The way the story has been set up, lots of people have taken this challenge before you. Those that have taken 1 box have walked away with a million dollars. Those that have taken both have walked away with a thousand dollars. You'd prefer to be in the first group to being in the second group, so you should take just one box.

The argument for taking both boxes is also easy. Either the demon has put the million in the opaque or she hasn't. If she has, you're better off taking both boxes. That way you'll get \$1,001,000 rather than \$1,000,000. If she has not, you're better off taking both boxes. That way you'll get \$1,000 rather than \$0. Either way, you're better off taking both boxes, so you should do that.

Both arguments seem quite strong. The problem is that they lead to incompatible conclusions. So which is correct?

6.1 Two Principles of Decision Theory

The puzzle was first introduced to philosophers by Robert Nozick, who in turn credited the puzzle to William Newcomb. And he suggested that the puzzle posed a challenge for the compatibility of two decision theoretic rules. These rules are

- Never choose dominated options
- Maximise expected utility

Nozick argued that if we never chose dominated options, we would choose both boxes. The reason for this is clear enough. If the demon has put \$1,000,000 in the opaque box, then it is better to take both boxes, since getting \$1,001,000 is better than getting \$1,000,000. And if the demon put nothing in the opaque box, then your choices are \$1,000 if you take both boxes, or \$0 if you take just the empty box. Either way, you're better off taking both boxes. This is obviously just the standard argument for taking both boxes. But note that however plausible it is as an argument for taking both boxes, it is compelling as an argument that taking both boxes is a dominating option.

To see why Nozick thought that maximising expected utility leads to taking one box, we need to see how he is thinking of the expected utility formula. That formula takes as an input the probability of each state. Nozick's way of approaching things, which was the standard at the time, was to take the expected utility of an action A to be given by the following sum

$$Exp(U(A)) = \Pr(S_1|A)U(AS_1) + ... + \Pr(S_n|A)U(AS_n)$$

Note in particular that we put into this formula the probability of each state *given that A is chosen*. We don't take the unconditional probability of being in that state. These numbers can come quite dramatically apart.

In Newcomb's problem, it is actually quite hard to say what the probability of each state is. (The states here, of course, are just that there is either \$1,000,000 in the opaque box or that there is nothing in it.) But what's easy to say is the probability of each state given the choices you make. If you choose both boxes, the probability that there is nothing in the opaque box is very high, and the probability that there is $1,000,000 in it is very low. Conversely, if you choose just the one box, the probability that there is $1,000,000 in it is very high, and the probability that there is nothing in it is very low. Simplifying just a little, we'll say that this high probability is 1, and the low probabiilty is 0. The expected utility of each choice then is given by the following formulae. (We'll write $C1$ for choosing one box, $C2$ for choosing both boxes, OM for million in opaque box, and ON for nothing in opaque box.

$$\begin{aligned}
&Exp(U(\text{Take both boxes}))\\
&\qquad = \Pr(OM|C2)U(C2 \wedge OM) + \Pr(ON|C2)U(ON \wedge C2)\\
&\qquad = 0 \times 1{,}001{,}000 + 1 \times 1{,}000\\
&\qquad = 1{,}000\\
&Exp(U(\text{Take one box}))\\
&\qquad = \Pr(OM|C1)U(C1 \wedge OM) + \Pr(ON|C1)U(ON \wedge C1)\\
&\qquad = 1 \times 1{,}000{,}000 + 0 \times 0\\
&\qquad = 1{,}000{,}000
\end{aligned}$$

I've assumed here that the marginal utility of money is constant, so we can measure utility by the size of the numerical prize. That's an idealisation, but hopefully a harmless enough one.

In earlier chapters we argued that the expected utility rule never led to a conflict with the dominance principle. But here it has led to a conflict. Something seems to have gone badly wrong. The problem was that we've used two distinct definitions of expected utility in the two arguments. In the version we had used in previous chapters, we presupposed that the probability of the states was independent of the choices that were made. So we didn't talk about $\Pr(S_1|A)$ or $\Pr(S_1|B)$ or whatever. We simply talked about $\Pr(S_1)$.

If you make that assumption, expected utility maximisation does indeed imply dominance. We won't rerun the entire proof here, but let's see how it works in this particular case. Let's say that the probability that there is $1,000,000 in the opaque box is x. It won't matter at all what x is. And assume that the expected utility of a choice A is given by this formula, where we use the unconditional probability of states as inputs.

$$Exp(U(A)) = \Pr(S_1)U(AS_1) + ... + \Pr(S_n|A)U(AS_n)$$

Applied to our particular case, that would give us the following calculations.

$$
\begin{aligned}
Exp(U(\text{Take both boxes})) & \\
&= \Pr(OM)U(C2 \wedge OM) + \Pr(ON)U(C2 \wedge ON) \\
&= x \times 1{,}001{,}000 + (1-x) \times 1{,}000 \\
&= 1{,}000 + 1{,}000{,}000x \\
Exp(U(\text{Take one box})) & \\
&= \Pr(OM)U(C1 \wedge OM) + \Pr(ON)U(C1 \wedge ON) \\
&= x \times 1{,}000{,}000 + (1-x) \times 0 \\
&= 1{,}000{,}000x
\end{aligned}
$$

And clearly the expected value of taking both boxes is 1,000 higher than the expected utility of taking just one box. So as long as we don't conditionalise on the act we are performing, there isn't a conflict between the dominance principle and expected utility maximisation.

While that does resolve the mathematical puzzle, it hardly resolves the underlying philosophical problem. Why, we might ask, shouldn't we conditionalise on the actions we are performing? In general, it's a bad idea to throw away information, and the choice that we're about to make is a piece of information. So we might think it should make a difference to the probabilities that we are using.

The best response to this argument, I think, is that it leads to the wrong results in Newcomb's problem, and related problems. But this is a somewhat controversial clam. After all, some people think that taking one box is the right result in Newcomb's problem. And as we saw above, if we conditionalise on our action, then the expected utility of taking one box is higher than the expected utility of taking both. So such theorists will not think that it gives the wrong answer at all. To address this worry, we need to look more closely back at Newcomb's original problem, and its variants.

6.2 Arguments for Taking Both Boxes

There are a couple of reasons for thinking we should take both boxes. One of these hypothesises a helpful friend, the other argues by analogy with some realistic cases.

Well Meaning Friends

The simplest argument is just a dramatisation of the dominance argument. But still, it is a way to see the force of that argument. Imagine that you have a friend who can see into the opaque box. Perhaps the box is clear from behind, and your friend is standing behind the box. Or perhaps your friend has super-powers that let them see into opaque boxes. If your friend was able to give you advice, and has your best interests at heart, they'll tell you to take both boxes. That's true whether or not there is a million dollars in the opaque box. Either way, they'll know that you're better off taking both boxes.

Of course, there are lots of cases where a friend with more knowledge than you and your interests at heart will give you advice that is different to what you might intuitively think is correct. Imagine that I have just tossed a biased coin that has an 80% chance of landing heads. The coin has landed, but neither of us can see how it has landed. I offer you a choice between a bet that pays $1 if it landed heads, and a bet that pays $1 if it landed tails. Since heads is more likely, it seems you should take the bet on heads. But if the coin has landed tails, then a well meaning and well informed friend will tell you that you should bet on tails.

But that case is somewhat different to the friend in Newcomb's problem. The point here is that you know what the friend will tell you. And plausibly, whenever you know what advice a friend will give you, you should follow that advice. Even in the coin-flip case, if you knew that your friend would tell you to bet on tails, it would be smart to bet on tails. After all, knowing that your friend would give you that advice would be equivalent to knowing that the coin landed tails. And if you knew the coin landed tails, then whatever arguments you could come up with concerning chances of landing tails would be irrelevant. It did land tails, so that's what you should bet on.

There is another way to dramatise the dominance argument. Imagine that after the boxes are opened, i.e. after you know which state you are in, you are given a chance to revise your choice if you pay $500. If you take just one box, then whatever is in the opaque box, this will be a worthwhile switch to make. It will either take you from $0 to $500, or from $1,000,000 to $1,000,500. And once the box is open, there isn't even an intuition that you should worry about how the box got filled. So you should make the switch.

But it seems plausible in general that if right now you've got a chance to do X, and you know that if you don't do X now you'll certainly pay good money to do X later, and you know that when you do that you'll be acting perfectly rationally, then you should simply do X. After all, you'll get the same result whether you do X now or later, you'll simply not have to pay the 'late fee' for taking X any later. More relevantly to our case, if you would switch to X once the facts were known, even if doing so required paying a fee, then it seems plausible that you should simply do X now. It doesn't seem that including the option of switching after the boxes are revealed changes anything about what you should do before the boxes are revealed, after all.

Ultimately, I'm not sure that either of the arguments I gave here, either the well meaning friend argument or the switching argument, are any more powerful than the dominance argument. Both of them are just ways of dramatising the dominance argument. And someone who thinks that you should take just one box is, by definition, someone who isn't moved by the dominance argument. In the next set of notes we'll look at other arguments for taking both boxes.

Real Life Newcomb Cases

It would be useful to come up with more realistic examples where the two approaches to utility theory, the one that uses conditional probabilities and the one that uses unconditional probabilities, come apart. It turns out that what is driving the divergence between the equations is that there is a common cause of the world being in a certain state and you making the choice that you make. Any time there is something in the world that tracks your decision making processes, we'll have a Newcomb like problem.

For example, imagine that we are in a Prisoners' Dilemma situation where we know that the other prisoner uses very similar decision making procedures to what we use. Here is the table for a Prisoners' Dilemma.

	Other Cooperates	Other Defects
You Cooperate	3,3	(0, 5
You Defect	5, 0	1, 1

In this table the notation x, y means that you get x utils and the other person gets y utils. Remember that utils are meant to be an overall measure of what you value, so it includes your altruistic care for the other person.

Let's see why this resembles a Newcomb problem. Assume that conditional on your performing an action A, the probability that the other person will do the same action is 0.9. Then, if we are taking probabilities to be conditional on choices, the expected utility of the two choices is

$$\begin{aligned} Exp(U(Coop)) &= 0.9 \times 3 + 0.1 \times 0 \\ &= 2.7 \\ Exp(U(Defect)) &= 0.1 \times 5 + 0.9 \times 1 \\ &= 1.4 \end{aligned}$$

So if we use probabilities conditional on choices, we end up with the result that you should cooperate. But note that cooperation is dominated by defection. If the other person defects, then your choice is to get 1 (by defecting) or 0 (by cooperating). You're better off cooperating. If the other person cooperates, then your choice is to get 5 (by defecting) or 0 (by cooperating). So whatever probability we give to the possible actions of the other person, provided we don't conditionalise on our choice, we'll end up deciding to defect.

Prisoners' Dilemma cases are much less fantastic than Newcomb problems. Even Prisoners' Dilemma cases where we have some confidence that the other party sufficiently resembles us that they will likely (not certainly) make the same choice as us are fairly realistic. So they are somewhat better than Newcomb's original problem for detecting intuitions. But the problem of divergent intuitions still remains. Many people are unsure about what the right thing to do in a Prisoners' Dilemma problem is. (We'll come back to this point when we look at game theory.)

So it is worth looking at some cases without that layer of complication. Real life cases are tricky to come by, but for a while some people suggested that the following might be a case. We've known for a long time that smoking causes various cancers. We've known for even longer than that that smoking is correlated with various cancers. For a while there was a hypothesis that smoking did not cause cancer, but was correlated with cancer because there was a common cause. Something, presumably genetic, caused people to (a) have a disposition to smoke, and (b) develop cancer. Crucially, this hypothesis went, smoking did not raise the risk of cancer; whether you got cancer or not was largely due to the genes that led to a desire for smoking.

We now know, by means of various tests, that this isn't true. (For one thing, the reduction in cancer rates among people who give up smoking is truly impressive, and hard to explain on the model that these cancers are all genetic.) But at least at some point in history it was a not entirely crazy hypothesis. Let's assume this hypothesis is actually true (contrary to fact). And let's assume that you (a) want to smoke, other things being equal, and (b) really don't want to get cancer. You don't know whether you have the desire for smoking/disposition to get cancer gene or not? What should you do?

Plausibly, you should smoke. You either have the gene or you don't. If you do, you'll probably get cancer, but you can either get cancer while smoking, or get cancer while not smoking, and since you enjoy smoking, you should smoke. If you don't, you won't get cancer whether you smoke or not, so you should indulge your preference for smoking.

It isn't just philosophers who think this way. At some points (after the smoking/cancer correlation was discovered but before the causal connection was established) various tobacco companies were trying very hard to get evidence for this 'common cause' hypothesis. Presumably the reason they were doing this was because they thought that if it were true, it would be rational for people to smoke more, and hence people would smoke more.

But note that this presumption is true if and only if we use the 'unconditional' version of expected utility theory. To see this, we'll use the following table for the various outcomes.

	Get Cancer	Don't get Cancer
Smoke	1	6
Don't Smoke	0	5

The assumption is that not getting cancer is worth 5 to you, while smoking is worth 1 to you. Now we know that smoking is evidence that you have the cancer gene, and this raises dramatically the chance of you getting cancer. So the (evidential) probability of getting cancer conditional on smoking is, we'll assume, 0.8, while the (evidential) probability of getting cancer conditional on not smoking is, we'll assume, 0.2. And remember this isn't because cancer causes smoking in our example, but rather that there is a common cause of the two. Still, this is enough to make the expected utilities work out as follows.

$$\begin{aligned} Exp(U(Smoke)) &= 0.8 \times 1 + 0.2 \times 6 \\ &= 2 \\ Exp(U(NoSmoke)) &= 0.2 \times 0 + 0.8 \times 5 \\ &= 4 \end{aligned}$$

And the recommendation is not to smoke, even though smoking dominates. This seems very odd. As it is sometimes put, the recommendation here seems to be a matter of managing the 'news', not managing the outcome. What's bad about smoking is that if you smoke you get some evidence that something bad is going to happen to you. In particular, you get evidence that you have this cancer gene, and that's really bad news to get because dramatically raises the probability of getting cancer. But not smoking doesn't mean that you don't have the gene, it just means that you don't find out that you have the gene. Not smoking looks like a policy of denying yourself good outcomes because you don't want to get bad news. And this doesn't look rational.

So this case has convinced a lot of decision theorists that we shouldn't use conditional probabilities of states when working out the utility of various outcomes. Using conditional probabilities will be good if we want to learn the 'news value' of some choices, but not if we want to learn how useful those choices will be to us.

6.3 Tickle Defence

Not everyone has been convinced by these 'real-life' examples. The counterargument is that in any realistic case, the gene that leads to smoking has to work by changing our dispositions. So there isn't just a direct causal connection between some genetic material and smoking. Rather, the gene causes a desire to smoke, and the desire to smoke cause the smoking. As it is sometimes put, between the gene and the smoking there has to be something mental, a 'tickle' that leads to the smoking.

Now this is important because we might think that rational agents know their own mental states. Let's assume that for now. So if an agent has the smoking desire they know it, perhaps because this desire has a distinctive phenomenology, a tickle of sorts. And if the agent knows this, then they won't get any extra evidence that they have a desire to smoke from their actual smoking. So the probability of getting cancer given smoking is not higher than the probability of getting cancer given not smoking.

In the case we have in mind, the bad news is probably already here. Once the agent realises that their values are given by the table above, they've already got the bad news. Someone who didn't have the gene wouldn't value smoking more than not smoking. Once the person conditionalises on the fact that that is their value table, the evidence that they actually smoke is no more evidence. Either way, they are (say) 80% likely to get cancer. So the calculations are really something like this:

$$\begin{aligned}
Exp(U(Smoke)) &= 0.8 \times 1 + 0.2 \times 6 \\
&= 2 \\
Exp(U(NoSmoke)) &= 0.8 \times 0 + 0.2 \times 5 \\
&= 1
\end{aligned}$$

And we get the correct answer that in this situation we should smoke. So this isn't a case where the two different equations we've used give different answers. And hence it isn't a reason for using unconditional probabilities rather than conditional probabilities.

There are two common responses to this argument. The first is that it isn't clear that there is always a 'tickle'. The second is that it isn't a requirement of rationality that we know what tickles we have. Let's look at these in turn.

First, it was crucial to this defence that the gene (or whatever) that causes both smoking and cancer causes smoking by causing some particular mental state first. But this isn't a necessary feature of the story. It might be that, say, everyone has the 'tickle' that goes along with wanting to smoke. (Perhaps this desire has some evolutionary advantage. Or, more likely, it might be a result of something that genuinely had evolutionary advantage.) Perhaps what the gene does is to affect how much willpower we have, and hence how likely we are to overcome the desire.

Second, it was also crucial to the defence that it is a requirement of rationality that people know what 'tickles' they have. If this isn't supposed, we can just imagine that our agent is a rational person who is ignorant of their own desires. But this supposition is quite strong. It is generally not a requirement of rationality that we know things about the external world. Some things are just hidden from us, and it isn't a requirement of rationality that we be able to see what is hidden. Similarly, it seems at least possible that some things in our own mind should be hidden. Whether or not you believe in things like subconscious desires, the possibility of them doesn't seem to systematically undermine human rationality.

Note that these two responses dovetail nicely. If we think that the gene works not by producing individual desires, but by modifying quite general standing dispositions like how much willpower we have, it is even more plausible to think that this is not something a rational person will always know about. It is a little odd to think of a person who desires to smoke but doesn't realise that they desire to smoke. It isn't anywhere near as odd to think about a person who has very little willpower but, perhaps because their willpower is rarely tested, doesn't realise that they have low willpower. Unless they are systematically ignoring evidence that they lack willpower, they aren't being clearly irrational.

So it seems there are possible, somewhat realistic, cases where one choice is evidence, to a rational agent, that something bad is likely to happen, even though the choice does not bring about the bad outcome. In such a case using conditional probabilities will lead to avoiding the bad news, rather than producing the best outcomes. And that seems to be irrational.

6.4 Causal and Evidential Decision Theory

So far in this chapter, we've looked at two ways of thinking about the expected utility of an action A. These are

$$\Pr(S_1)U(S_1A) + ... + \Pr(S_n)U(S_nA)$$
$$\Pr(S_1|A)U(S_1A) + ... + \Pr(S_n|A)U(S_nA)$$

It will be convenient to have names for these two approaches. So let's say that the first of these, which uses unconditional probabilities, is **causal expected value**, and the second of these, which uses conditional probabilities is the **evidential expected value**. The reason for the names should be clear enough. The causal expected value measures what you can expect to bring about by your action. The evidential expected value measures what kind of result your action is evidence that you'll get.

Causal Decision Theory then is the theory that rational agents aim to maximise causal expected utility.

Evidential Decision Theory is the theory that rational agents aim to maximise evidential expected utility.

Over the past two chapters we've been looking at reasons why we should be causal decision theorists rather than evidential decision theorists. We'll close out this chapter by looking at various puzzles for causal decision theory, and then looking at one reason why we might want some kind of hybrid approach.

Right and Wrong Tabulations

If we use the causal approach, it is very important how we divide up the states. We can see this by thinking again about an example from Jim Joyce that we discussed a while ago.

> Suupose you have just parked in a seedy neighborhood when a man approaches and offers to "protect" your car from harm for $10. You recognize this as extortion and have heard that people who refuse "protection" invariably return to find their windshields smashed. Those who pay find their cars intact. You cannot park anywhere else because you are late for an important meeting. It costs $400 to replace a windshield. Should you buy "protection"? Dominance says that you should not. Since you would rather have the extra $10 both in the even that your windshield is smashed and in the event that it is not, Dominance tells you not to pay. (Joyce, *The Foundations of Causal Decision Theory*, pp 115-6.)

If we set this up as a table, we get the following possible states and outcomes.

	Broken Windshield	Unbroken Windshield
Pay extortion	-$410	-$10
Don't pay	-$400	0

Now if you look at the causal expected value of each action, the expected value of not paying will be higher. And this will be so whatever probabilities you assign to broken windshield and unbroken windshield. Say that the probability of the first is x and of the second is $1 - x$. Then we'll have the following (assuming dollars equal utils)

$$\begin{aligned} Exp(U(\text{Pay extortion})) &= -410x - 10(1-x) \\ &= -400x - 10 \\ Exp(U(\text{Don't pay}) &= -400x - 0(1-x) \\ &= -400x \end{aligned}$$

Whatever x is, the causal expected value of not paying is higher by 10. That's obviously a bad result. Is it a problem for causal decision theory though? No. As the name 'causal' suggests, it is crucial to causal decision theory that we separate out what we have causal power over from what we don't have causal power over. The states of the world represent what we can't control. If something can be causally affected by our actions, it can't be a background state.

So this is a complication in applying causal decision theory. Note that it is not a problem for evidential decision theory. We can even use the very table that we have there. Let's assume that the probability of broken windshield given paying is 0, and the probability of unbroken windshield given paying is 0. Then the expected utilities will work out as follows

$$\begin{aligned} Exp(U(\text{Pay extortion})) &= -410 \times 0 - 10 \times 1 \\ &= -10 \\ Exp(U(\text{Don't pay}) &= -400 \times 1 - 10 \times 0 \\ &= -400 \end{aligned}$$

So we get the right result that we should pay up. It is a nice feature of evidential decision theory that we don't have to be so careful about what states are and aren't under our control. Of course, if the only reason we don't have to worry about what is and isn't under our control is that the theory systematically ignores such facts, even though they are intuitively relevant to decision theory, this isn't perhaps the best advertisement for evidential decision theory.

Why Ain'Cha Rich

There is one other argument for evidential decision theory that we haven't yet addressed. Causal decision theory recommends taking two boxes in Newcomb's problem; evidential decision theory recommends only taking one. People who take both boxes tend, as a rule, to end up poorer than people who take just the one box. Since the aim here is to get the best outcome, this might be thought to be embarrassing for causal decision theorists.

Causal decision theorists have a couple of responses to this argument. One response is to note that it seems to licence a perverse action in a variant on Newcomb's Problem. Change the case so that *both* boxes are transparent, but everything else is the same. So you have a choice of one box or two, the demon has put $1,000,000 in one of the boxes if and only if you'll take that box, and the demon has put $1,000 in the other box. What do you do? Presumably, when you can see the money, you take both boxes. But the people who take just one box walk away richer. So the "Why Ain'Cha Rich" argument seems to lead to the wrong answer.

Perhaps that response is too quick. Perhaps, you think, even when both boxes are transparent, you should just take the one box. That will be a reason for you not liking causal decision theory. But note that it is also a reason to not like evidential decision theory. Once you start following "Why Ain'Cha Rich" reasoning, you have to give up both causal and evidential decision theory.

Causal decision theorists have another response to this objection. They say that Newcomb's problem is a situation where there is someone who is quite smart, and quite determined to reward irrationality. In such a case, they say, it isn't too surprising that irrational people, i.e. evidential decision theorists, get rewarded. Moreover, if a rational person like them were to have taken just one box, they would have ended up with even less money, i.e., they would have ended up with nothing. But this move is also controversial. One could argue that it is rational to adjust one's behaviour to the demon's expected actions, so there is no such thing as rewarding irrationality. This seems to be a standoff.

Dilemmas

Consider the following story, told by Allan Gibbard and William Harper in their paper setting out causal decision theory.

> Consider the story of the man who met Death in Damascus. Death looked surprised, but then recovered his ghastly composure and said, 'I AM COMING FOR YOU TOMORROW'. The terrified man that night bought a camel and rode to Aleppo. The next day, Death knocked on the door of the room where he was hiding, and said 'I HAVE COME FOR YOU'.
>
> 'But I thought you would be looking for me in Damascus', said the man.
>
> 'NOT AT ALL', said Death 'THAT IS WHY I WAS SURPRISED TO SEE YOU YESTERDAY. I KNEW THAT TODAY I WAS TO FIND YOU IN ALEPPO'.
>
> Now suppose the man knows the following. Death works from an appointment book which states time and place; a person dies if and

> only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with Death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo...
>
> If... he decides to go to Aleppo, he then has strong grounds for expecting that Aleppo is where Death already expects him to be, and hence it is rational for him to prefer staying in Damascus. Similarly, deciding to stay in Damascus would give him strong grounds for thinking that he ought to go to Aleppo.

In cases like this, the agent is in a real dilemma. Whatever he does, it seems that it will be the wrong thing. If he goes to Aleppo, then Death will probably be there. And if he stays in Damascus, then Death will probably be there as well. So it seems like he is stuck.

Of course in one sense, there is clearly a right thing to do, namely go wherever Death isn't. But that isn't the sense of right decision we're typically using in decision theory. Is there something that he can do that maximises expected utility. In a sense the answer is "No". Whatever he does, doing that will be some evidence that Death is elsewhere. And what he should do is go wherever his evidence suggests Death isn't. This turns out to be impossible, so the agent is bound not to do the rational thing.

Is this a problem for causal decision theory? It is if you think that we should always have a rational option available to us. If you think that 'rational' here is a kind of 'ought', and you think 'ought' implies 'can', then you might think we have a problem, because in this case there's a sense in which the man can't do the right thing. (Though this is a bit unclear; in the actual story, there's a perfectly good sense in which he could have stayed in Aleppo, and the right thing to do, given his evidence, would have been to stay in Aleppo. So in one sense he could have done the right thing.) But both the premises of the little argument here are somewhat contentious. It isn't clear that we should say you ought, in any sense, maximise expected utility. And the principle that ought implies can is rather controversial. So perhaps this isn't a clear counterexample to causal decision theory.

A different objection to this response is that saying "It's a dilemma" implies that the two options are symmetric. And while that's a plausible thing to say in this case, it is less plausible in some other cases. Reed Richter (in a 1984 *Australasian Journal of Philosophy* paper) argued that this was the wrong thing to say about *asymmetric* versions of the Death in Damascus case. Imagine that getting to Aleppo will cost a huge amount of money, and be incredibly painful. Then the table might look something like this:

	Death in Damascus	Death in Aleppo
Go to Damascus	-1	0.5
Go to Aleppo	1	-1.5

Again, whatever the man does, he will regret it, just like in the original Death in Damascus example. But it seems wrong to treat the two options available to the man symmetrically. After all, going to Aleppo is much worse for him. If forced to choose, some have argued, he should stay in Damascus. Causal decision theory doesn't generate that answer, so if you think it is true, you think there is a problem her for causal decision theory.

6.5 Weak Newcomb Problems

Imagine a small change to the original Newcomb problem. Instead of there being $1000 in the clear box, there is $800,000. Still, evidential decision theory recommends taking one box. The evidential expected value of taking both boxes is now roughly $800,000, while the evidential expected value of taking just the one box is $1,000,000. Causal decision theory recommends taking both boxes, as before.

So neither theory changes its recommendations when we increase the amount in the clear box. But I think many people find the case for taking just the one box to be less compelling in this variant. Does that suggest we need a third theory, other than just causal or evidential decision theory?

It turns out that we can come up with hybrid theories that recommend taking one box in the original case, but two boxes in the original case. Remember that in principle anything can have a probability, including theories of decision. So let's pretend that given the (philosophical) evidence on the table, the probability of causal decision theory is, say, 0.8, while the probability of evidential decision theory is 0.2. (I'm not saying these numbers are right, this is just a possibility to float.) And let's say that we should do the thing that has the highest *expected* expected utility, where we work out expected expected utilities by summing over the expectation of the action on different theories, times the probability of each theory. (Again, I'm not endorsing this, just floating it.)

Now in the original Newcomb problem, evidential decision theory says taking one boxes is $999,000 better, while causal decision theory say staking both boxes is $1,000 better. So the expected expected utility of taking one box rather than both boxes is $0.2 \times 999,000 - 0.8 \times 1,000$, which is 199,000. So taking one box is 'better' by 199,000

In the modified Newcomb problem, evidential decision theory says taking one boxes is \$200,000 better, while causal decision theory says taking both boxes is \$800,000 better. So the expected expected utility of taking one box rather than both boxes is $0.2 \times 200,000 - 0.8 \times 800,000$, i.e., -600,000. So taking both boxes is 'better' by 600,000.

If you think that changing the amount in the clear box can change your decision in Newcomb's problem, then possibly you want a hybrid theory, perhaps like the one floated here.

Chapter 7

Introduction to Games

Game theory is a slighttly oddly defined subject matter. A **game** is any decision problem where the outcome depends on the actions of more than one agent, as well as perhaps on other facts about the world. **Game Theory** is the study of what rational agents do in such situations. You might think that the way to figure that out would be to come up with a theory of how rational agents solve decision problems, i.e., figure out **Decision Theory**, and then apply it to the special case where the decision problem involves uncertainty about the behaviour of other rational agents. And I think that's broadly correct.

But as we'll see a little as we go along, there are useful reasons for looking at games differently from how we look at decision problems. One of those reasons is historical; the theory of games has developed somewhat independently of the theory of decisions, and so it is useful to study game theory independently to follow these historical trends. The other reason concerns epistemology. Most of the time, we can't solve a decision problem unless we are given some probabilities to use as inputs to our expected utility calculations. But it often turns out that we can solve a game without such an input, but with only the assumption that the other player is rational.

Let's start with a very simple example of a game. Each player in the game is to choose a letter, A or B. After they make the choice, the player will be paired with a randomly chosen individual who has been faced with the very same choice. They will then get rewarded according to the following table.

- If they both choose A, they will both get $1
- If they both choose B, they will both get $3
- If one chooses A, and the other B, the one who chose A will get $5, and the one who chose B will get $0.

We can represent this information in a small table, as follows. (Where possible, we'll use uppercase letters for the choices on the rows, and lowercase letters for choices on the columns.)

	Choose a	Choose b
Choose A	\$1, \$1	\$5, \$0
Choose B	\$0, \$5	\$3, \$3

We represent one player, imaginatively called **Row**, or *R*, on the rows, and the other player, imaginatively called **Column**, or *C* on the columns. A cell of the table represents the outcomes, if *R* chose to be on that row, and *C* chose to be in that column. There are two monetary sums listed in each cell. We will put the row player's outcome first, and the column player's outcome second. You should verify that the table represents the same things as the text.

Now let's note a few distinctive features of this game.

- Whatever *C* does, *R* gets more money by choosing A. If *C* chooses a, then *R* gets \$1 if she chooses A, and \$0 if she chooses B; i.e., she gets more if she chooses A. And if *C* chooses b, then *R* gets \$5 if she chooses A, and \$3 if she chooses B; i.e., she gets more if she chooses A.
- Since the game is symmetric, that's true for *C* as well. Whatever *R* does, she gets more money if she chooses a.
- But the players collectively get the most money if they both choose B.

So doing what maximises the players' individual monetary rewards does not maximise, indeed it minimises, their collective monetary rewards.

I've been careful so far to distinguish two things: the monetary rewards each player gets, and what is best for each player. More elegantly, we need to distinguish the **outcomes** of a game from the **payoffs** of a game. The outcomes of the game are things we can easily physically describe: this player gets that much money, that player goes to jail, this other player becomes President of a failing Middle Eastern dictatorship, etc. The payoffs of a game describe how well off each player is with such an outcome. Without knowing much about the background of the players, we don't know much about the payoffs. We've already seen a variant of this point, when we considered the distinction between outcomes in dollars and payoffs in utils. It is because these come apart that insurance contracts can be sensible investments for both parties. But the difference can be even more dramatic in the case of games.

For instance, the following two games are both consistent with the above description of the outcomes. In the first game, each player cares only about their own monetary rewards, so the number of utils they get equals the number of dollars they get. In the second, everyone cares about how much money the group consisting of the two players gets, and not at all about how much of the group's money they get. Perhaps they care about this because they are altruistic, or because they know the proceeds will be shared; it doesn't really matter for present purposes. What does matter is that the individual outcomes are a function of the payoff to the group.

Game 1	Choose a	Choose b
Choose A	1, 1	5, 0
Choose B	0, 5	3, 3

Game 2	Choose a	Choose b
Choose A	2, 2	5, 5
Choose B	5, 5	6, 6

You might note that I've started numbering the games, and that I didn't number the initial description of the outcomes. There's a reason for this. Technically, we'll say that a game is specified by setting out what moves, or as we'll sometimes call them, *strategies* are available for each player, and what the *payoffs* are for each player, given the moves that they make. (And, perhaps, the state of the world; for now we're just looking at games where only moves matter for payoffs. And we're only looking for now at games where each player makes a simultaneous choice of strategy. We'll return to how general an account this is in a little while.) Specifying the outcome of a game in physical terms doesn't give us a unique game. We need to know more to get a genuine game specification.

7.1 Prisoners' Dilemma

Game 1 is often called a **Prisoners' Dilemma**. There is perhaps some terminological confusion on this point, with some people using the term "Prisoners' Dilemma" to pick out any game whose *outcomes* are like those in the games we've seen so far, and some using it only to pick out games whose *payoffs* are like those in Game 1. Following what Simon Blackburn says in "Practical Tortoise Raising", I think it's not helpful to use the the term in the first way. So I'll only use it for games whose payoffs are like those in Game 1. And what I mean by payoffs like those in Game 1 is the following pair of features.

- Each player is better off choosing A than B, no matter what the other player does.

- The players would both be better off if they both chose B rather than both chose A.

You might want to add a third condition, namely that the payoffs are symmetric. But just what that could *mean* is a little tricky. It's easy to compare *outcomes* of different players; it's much harder to compare *payoffs*. So we'll just leave it with these two conditions.

It is often very bad to have people in a Prisoners' Dilemma situation; everyone would be better off if they were out of it. Or so it might seem at first. Actually, what's really true is that the two players would be better off if they were out of the Prisoners' Dilemma situation. Third parties might stand to gain quite a lot from it. (If I'm paying out the money at the end of the game, I prefer that the players are in Game 1 to Game 2.) We'll come back to this point in a little. There are several ways we could try and escape a Prisoners' Dilemma. We'll mention four here, the first two of which we might naturally associate with Adam Smith.

The first way out is through **compassion**. If each of the players cares exactly as much about the welfare of the other player as they do about themselves, then we'll be in something like Game 2, not Game 1. Note though that there's a limit to how successful this method will be. There are variants of the Prisoners' Dilemma with arbitrarily many players, not just two. In these games, each player is better off if they choose A rather than B, no matter what the others do, but all players are better off if all players choose B rather than A. It stretches the limit of compassion to think we can in practice value each of these players's welfar equally to our own.

Moreover, even in the two player game, we need exact match of interests to avoid the possibility of a Prisoners' Dilemma. Let's say that *R* and *C* care about each other's welfare a large amount. In any game they play for money, each players' payoff is given by the number of pounds that player wins, plus 90% of the number of pounds the other player wins. Now let's assume they play a game with the following outcome structure.

	Choose a	Choose b
Choose A	$9.50, $9.50	$20, $0
Choose B	$0, $20	$10, $10

So we'll have the following payoff matrix.

Game 3	Choose a	Choose b
Choose A	18.05, 18.05	20, 18
Choose B	18, 20	19, 19

And that's still a Prisoners' Dilemma, even though the agents are very compassionate. So compassion can't do all the work. But probably none of the other 'solutions' can work unless compassion does some of the work. (That's partially why Adam Smith wrote the *Theory of Moral Sentiments* before going on to economic work; some moral sentiments are necessary for economic approaches to work.)

Our second way out is through **contract**. Let's say each party contracts with the other to choose B, and agrees to pay $2.50 to the other if they break the contract. Assuming that this contract will be enforced (and that the parties know this), here is what the outcome table now looks like.

	Choose a	Choose b
Choose A	$1, $1	$2.50, $2.50
Choose B	$2.50, $2	$3, $3

Now if we assume that the players just value money, those outcomes generate the following game.

Game 4	Choose a	Choose b
Choose A	1,1	2.5, 2.5
Choose B	2.5, 2.5	3,3

Interestingly, the game looks just like the original Prisoners' Dilemma as played between members of a commune. Basically, the existence of side contracts is enough to turn capitalists into communists.

A very closely related approach, one which is typically more efficient in games involving larger numbers of players, is to modify the outcomes, and hence the payoffs, with taxes. A striking modern example of this involves congestion charges in large cities. There are many circumstances where each person would prefer to drive somewhere than not, but if everyone drives, we're all worse off than if everyone took mass transit (or simply stayed home). The natural solution to this problem is simply to put a price on driving into the congested area. If the price is set at the right level, those who pay the charge are better off than if the charge was not there, since the amount they lose through the charge is gained back through the time they save.

In principle, we could always avoid Prisoners' Dilemma situations from arising through judicious use of taxes and charges. But it's hard to get the numbers right, and even harder to do the enforcement. So sometimes states will try to solve Prisoners' Dilemma situations with **regulation**. We see this in Beijing, for example, when they try to deal with congestion not by charging people money to enter the city, but by simply banning (certain classes of) people from driving into the city on given days. At a more abstract level, you might think of ethical prohibitions on 'free-riding' as being ways of morally regulating away certain options. If choosing B is simply ruled out, either by law or morality, there's clearly no Prisoners' Dilemma!

Having said that, the most important kind of regulation around here concerns making sure Prisoners' Dilemma situations survive, and are not contracted away. Let the two players be two firms in a duopoly; i.e., they are the only firms to provide a certain product. It is common for there to be only two firms in industries that require massive capital costs to startup, e.g., telecommunications or transport. In small towns, it is common to have only two firms in more or less every sphere of economic life. In such cases there will usually be a big distance between the prices consumers are prepared to pay for the product, and the lowest price that the firm could provide the product and still turn a profit. Call these prices High and Low.

If the firms only care about maximising profit, then it looks like setting prices to High is like choosing B in Game 1, and setting prices to Low is like choosing A in that game. The two firms would be better off if each of them had High prices. But if one had High prices, the other would do better by undercutting them, and capturing (almost) all the market. And if both had Low prices, neither would be better off raising prices, because (almost) everyone would desert their company. So the firms face a Prisoners' Dilemma.

As Adam Smith observed, the usual way businesses deal with this is by agreeing to raise prices. More precisely, he says,

> People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices.

And that's not too surprising. There's a state where they are both better off than the state where they can compete. If by changing some of the payoffs they can make that state more likely to occur, then they will. And that's something that we should regulate away, if we want the benefits of market competition to accrue to consumers.

The final way to deal with a Prisoners' Dilemma is through **iteration**. But that's a big, complicated issue, and one that we'll come back to much later in these notes.

7.2 Dominating Strategies

It's important to be careful about what we mean by a dominated strategy. Here is a more careful definition.

Strong Domination A strategy s_1 *strongly dominates* strategy s_2 for player i iff for any combination of moves by other players, and states of the external world, playing s_1 provides a greater payoff than playing s_2, assuming other players make those moves, and the world is that way.

Strongly Dominated A strategy is strongly dominated iff some other strategy, available to the same player, strongly dominates it.

There is a potential scope ambiguity in the description of a strongly dominated strategy that it is important to be clear about. The claim is *not* that a strategy is strongly dominated if no matter what else happens, some strategy or other does better than it. It is that a strategy is dominated if some particular strategy does better in every circumstance. We can see the difference between these two ideas in the following game.

Game 5	l	r
U	3, 0	0, 0
M	2, 0	2, 0
D	0, 0	3, 0

Consider this game from *R*'s perspective; who is choosing as always the rows. Her options are **U**p, **M**iddle and **D**own. *C* is choosing the columns; her choices are **l**eft or **r**ight. (I hope the ambiguity between *r* for *Right* and *R* for *Row* is not too confusing. It should be very hard in any given context to get them mixed up, and hopefully the convention we've adopted about cases will help.)

Notice that Middle is never the best outcome for *R*. If *C* chooses Left, *R* does best choosing Up. If *C* chooses Right, *R* does best choosing Down. But that does not mean Middle is dominated. Middle would only be dominated if one particular choice was better than it in both circumstances. And that's not true. Middle does better than Up in one circumstance (when *C* chooses Right) and does better than Down in another circumstance (when *C* chooses Left).

Indeed, there are situations where Middle might be uniquely rational. We need to say a bit more about expected utility theory to say this precisely, but consider what happens when *R* suspcts *C* is just going to flip a coin, and choose Left if it comes up Heads, and Right if it comes up Tails. (Since literally nothing is at stake for C in the game, this might be a reasonable hypothesis about what C will do.) Then it maximises R's **expected** return to choose Middle.

So far we've talked about the notion of strong dominance. We also need a notion of **weak dominance**. Roughly, strategy s_1 weakly dominates strategy s_2 if s_1 can do better than s_2, and can't do worse. More formally,

Weak Domination A strategy s_1 *weak dominates* strategy s_2 for player i iff for some combination of moves by other players, and states of the external world, playing s_1 provides a greater payoff than playing s_2, assuming other players make those moves, and the world is that way, and for all combination of moves by other players, and states of the external world, playing s_1 provides at least as high a payoff as playing s_2, assuming other players make those moves, and the world is that way,

Weakly Dominated A strategy is weakly dominated iff some other strategy, available to the same player, weakly dominates it.

It does seem plausible that agents should prefer any strategy over an alternative that it weakly dominates. This leads to distinctive results in games like the following.

Game 6	a	b
A	1, 1	0, 0
B	0, 0	0, 0

In this game, choosing A does not strongly dominate choosing B for either player. The game is symmetric, so from now on we'll just analyse it from R's perspective. The reason choosing A does not strongly dominate is is that if C chooses b, then choosing A leads to no advantage. R gets 0 either way.

But choosing A does *weakly* dominate choosing B. A does better than B in one circumstance, namely when C chooses a, and never does worse. So a player who shuns weakly dominated options will always choose A rather than B in this game.

7.3 Iterated Dominance

A rational player, we've argued, won't choose dominated strategies. Now let's assume, as is often the case, that we're playing a game where each player knows that the other player is rational. In that case, the players will not only decline to play dominated strategies, they will decline to play strategies that only produce the best outcomes if the other player adopts a dominated strategy. This can be used to generate a prediction about what people will, or at least should, do in various game. We can see this in the following game.

Game 7	Choose a	Choose b
Choose A	1, 2	5, 0
Choose B	0, 0	3, 6

If we look at things from C's perspective, neither strategy is dominated. She wants to choose whatever R chooses. But if we look at things from R's perspective, things are a little different. Here there is a strongly dominating strategy, namely choosing A. So C should really think to herself that there's no way R, who is rational, is going to choose B. Given that, the table really looks like this.

Game 7′	Choose a	Choose b
Choose A	1, 2	5, 0

I've put the prime there to indicate it is officially a different game. But really all I've done is delete a dominated strategy that the other player has. Now it is clear what C should do. In this 'reduced' game, the one with the dominated strategy deleted, there is a dominant strategy for C. It is choosing a. So C should choose a.

The reasoning here might have been a little convoluted, but the underlying idea is easy enough to express. R is better off choosing A, so she will. C wants to choose whatever R chooses. So C will choose a as well.

Let's go through a small variant of this game which might, after redescription, look fairly familiar.

Game 8	l	r
U	1, 1	1001, 0
D	0, 0	1000, 1

Just as in Game 7, R has a dominant strategy. It is to choose Up. (Again, I'm labelling just using the first letter of the description of the move.) And given that R will choose Up, the best thing for C to do is choose l. So it looks like we should end up in the top-left corner of the table, just like in Game 7.

Those of you who have taken some decision theory should recognise Game 8. It is just Newcomb's problem, with some assumptions about the payoffs. (If you don't know what Newcomb's Problem is, skip the next couple of paragraphs. We'll discuss Newcomb's problem in more detail later in the notes.) R in this case is the human player, who is usually the focus of attention in decision theory classes. Her payoffs here are just her payments in the usual statement of the game, divided by 1000. Up is her choosing both boxes, and Down is her choosing one box.

C is the demon. The demon isn't usually treated as a player in decision theoretic versions of the puzzle, but she clearly has views, and preferences. The demon wants to predict the move that the player makes. So we've represented her payoffs that way. Left is her predicting two boxes, Right is her predicting one box. And if she gets the prediction right, she gets a payoff of 1, if she gets it wrong, she gets a payoff of 0.

So Newcomb's problem is just a simple game, and it can be solved by noting that one player has a dominating strategy, and the other player, i.e., the demon, has a dominating strategy under the assumption that this dominating strategy is played.

We can use the idea of removing dominating strategies to illustrate some puzzling features of a couple of other games. I won't do tables for these games, because they'd be much too big. The first is a location game that has many applications.

> **Game 9**
>
> Two trucks have to choose where they will sell ice-cream on a particular beach. There are 11 locations to choose from, which we'll number 0, 1, ..., 9, 10. Spot 0 is at the left end of the beach, Spot 10 is at the right end of the beach, and the other spots are equally spaced in between. There are 10 people at each location. Each of them will buy ice-cream. If one truck is closer, they will buy ice-cream from that truck. If two trucks are equally close, then 5 of them will buy ice-cream from one truck, and 5 from the other. Each truck aims to maximise the amount of ice-cream it sells. Where should the trucks end up?

Let's start by looking at a fragment of the payoff matrix. The payoffs are numbers of ice-creams sold. We'll call the players *R* for Row and *C* for column, as usual, and just use the number *n* for the strategy of choosing location *n*.

	0	1	2	3	4	...
0	55, 55	10, 100	15, 95	20, 90	25, 85	...
1	100, 10	55, 55	20, 90	25, 95	30, 80	...
...						

I'll leave it as an exercise to confirm that these numbers are indeed correct. But there's something already from the numbers that we can see. No matter what C selects, R is better off picking 1 than 0. If C picks 0 or 1, she is a lot better off; she sells 45 more ice-creams. And if C picks a higher number, she is a bit better off; she sells 5 more ice-creams. So picking 1 dominates picking 2.

Let's look at the opposite corner of the matrix.

	...	6	7	8	9	10
...						
9	...	30, 80	25, 85	20, 90	55, 55	100, 10
10	...	25, 85	20, 80	15, 95	10, 100	55, 55

Again, there should be a pattern. No matter what C does, R is better off picking 9 than 10. In most cases, this leads to selling 5 more ice-creams. If C also picks 9 or 10, the picking 9 gives R a big advantage. The argument here was obviously symmetric to the argument about picking 0 or picking 1, so I'll stop concentrating on what happens when both players select high numbers, and focus from here on the low numbers.

So there is a clear conclusion to be drawn from what we've said so far.

- Spot 0 is dominated by Spot 1, and Spot 10 is dominated by Spot 9. So if R is rational, she won't pick either of those spots. Since the game is symmetric, if C is rational, she won't pick either of those spots either.

Now let's turn to the comparison between Spot 1 and Spot 2. Again, we'll just look at a fragment of the matrix.

	0	1	2	3	4	5	...
1	100, 10	55, 55	20, 90	25, 95	30, 80	35, 75	...
2	95, 15	90, 20	55, 55	30, 80	35, 75	40, 70	...
...							

The pattern is also clear. If C selects any number above 2, then R sells 5 more ice-creams by picking 2 rather than 1. If C selects either 1 or 2, then R sells 35 more ice-creams by picking 2 rather than 1. But if C selects 0, then R sells 5 *fewer* ice-creams by picking 2 rather than 1. So picking 2 does *not* dominate picking 1.

But note the only circumstance when picking 2 is worse than picking 1 is if C picks 0. And picking 0 is, for C, a strongly dominated strategy. So picking 2 is sure to do better than picking 1 if C does not play a strongly dominated strategy. Assuming C is rational, we can represent this by deleting from the game matrix

C's dominated options. Here's what the top left corner of the game matrix looks like when we do that.

	1	2	3	4	5	…
1	55, 55	20, 90	25, 95	30, 80	35, 75	…
2	90, 20	55, 55	30, 80	35, 75	40, 70	…
…						

And now it looks like picking 1 is dominated. This teaches us another lesson about the game.

- If we 'delete' the option of picking either 0 or 10 for *C*, then picking 2 dominates picking 1 for *R*. In other words, if *R* knows *C* is rational, and hence won't pick a dominated option, then picking 2 dominates picking 1 for *R*, relative to the space of epistemically possible moves in the game. For similar reasons, if *R* knows *C* is rational, then picking 8 dominates picking 9. And if *C* knows *R* is rational, then picking either 1 or 9 is dominated (by 2 and 8 respectively).

Summarising what we know so far,

- If the players are rational, they won't pick 0 or 10.
- If the players know the other player is rational, they also won't pick 1 or 9.

Let's continue down the matrix. For simplicity, we'll leave off columns 0 and 10, since they are dominated, and we have deleted those as possible options.

	1	2	3	4	5	…
2	90, 20	55, 55	30, 80	35, 75	40, 70	…
3	85, 25	80, 30	55, 55	40, 70	45, 65	…
…						

Picking 3 doesn't *quite* dominate picking 2. In most circumstances, *R* does better by picking 3 rather than 2. But she does a little worse if *C* picks 1. But wait! We just had an argument that *C* shouldn't pick 1. Or, at least, if *C* knows that *R* is rational, she shouldn't pick 1. Let's assume that *C* does know that *R* is rational, and *R* in turn knows that fact, so she can use it in reasoning. That means she knows *C* won't pick 1. So she can delete it from consideration too. And once she does, picking 3 dominates picking 2, relative to the reduced space of epistemically possible outcomes.

I haven't belaboured the point as much as in the previous paragraphs, but hopefully the following conclusion is clear.

- If the players know that the players know that the other player is rational, they won't pick 2 or 8.

And you can see where this is going. Once we rule out each player picking 2 or 8, then picking 4 dominates picking 3, so picking 3 should be ruled out. And picking 7 should be ruled out for symmetric reasons. But once 3 and 7 are ruled out, picking 5 dominates picking either 4 or 6. So both players will end up picking 5.

And that is, in the standard economic textbooks, a nice explanation of why we see so much 'clustering' of shops. (For instance, why so often there are several petrol stations on one corner, rather than spread over town.) Of course the full explanation of clustering is more common, but it is nice to see such a simple model deliver such a strong outcome.

This process is called the **Iterative Deletion of Dominated Strategies**. More precisely, what we've used here is the strategy of iteratively deleting **strongly** dominated strategies. This is a powerful technique for solving games that don't, at first glance, admit of any easy solution.

7.4 Iteration and Knowledge

It is worth getting clear on just how strong one's assumptions need to be to justify iteratively deleting dominated strategies. The standard terminology here can be a little confusing. After all, it isn't that picking 4 *dominates*, in any sense, picking 3. What's really true is that if we quantify over a restricted range of choices for C, then picking 4 is better for R than picking 3, no matter which choice *from that range*, C chooses. And that's a good reason to pick 4 rather than 3, provided that R knows that C will make a pick in that range. From that perspective, it's instructive to complete the list of lessons that we were compiling about the game.

- If the players are rational, they won't pick 0 or 10.
- If the players know the other player is rational, they also won't pick 1 or 9.
- If the players know that the players know that the other player is rational, they also won't pick 2 or 8.
- If the players know that the players know that the players know that the other player is rational, they won't pick 3 or 7.
- If the players know that the players know that the players know that the players know that the other player is rational, they also won't pick 4 or 6, i.e., they will pick 5

There are a lot of assumptions built in to all of this. It would be nice to have a way of summarising them. The standard approach traces back to David Lewis's *Convention*.

Common Knowledge In a game, it is common knowledge that *p* if each player knows it, each player knows that each player knows it, each player knows that each player knows that each player know it, and so on.

In many, but not all, games, we assume common knowledge of the rationality of the players. In Game 9, common knowledge of rationality makes picking 5 rationally mandatory.

There is a fun story that is usually told to illustrate the importance of common knowledge.

> **Slapville**
>
> In Slapville, it is culturally required to slap oneself if one is in public with a dirty face. Larry, Curly and Moe are in a room together, fortunately one without mirrors. Each of them has a dirty face, but they can't see their own faces, they can only see the other faces. And each face is dirty. Inspector Renault walks into the room and says, "I'm shocked! Someone in this room has a dirty face." After a long delay, Larry, Curly and Moe each slap themselves in the face (thereby getting dirty hands as well as dirty faces). Why?

One way to be puzzled by Slapville is to start with the concept of **mutual knowledge**. It is mutual knowledge that *p* if everyone in the game knows that *p*. In Slapville, it is mutual knowledge that someone has a dirty face. It is even, modulo Williamsonian concerns, mutual knowledge* that someone has a dirty face. (By *S* knows* that *p*, I mean *S* knows that *p*, and *S* knows that *S* knows *p*, and *S* knows that *S* knows that *S* knows that *p*, and so on.) So you might wonder what difference Renault's statement makes. After all, just like his namesake, he's just voicing something everyone already knew.

But it wasn't common knowledge that someone has a dirty face. Consider things from Larry's perspective. He knows someone has a dirty face. He can see Curly and Moe's dirty faces. And he knows that everyone knows that someone has a dirty face. He clearly knows it; he can see Curly and Moe. And Curly knows it; he can see Moe. And Moe knows it; he can see Curly.

But he doesn't know that everyone knows that everyone knows that someone has a dirty face. For all he knows, only Curly and Moe have dirty faces. If that's true, the only dirty face Curly knows about is Moe's. So for all Larry knows that Curly knows, only Moe has a dirty face. And if only Moe has a dirty face, then Moe doesn't know that someone has a dirty face. So for all Larry knows that Curly knows, Moe doesn't know that someone has a dirty face.

Or at least, that's the situation before Renault speaks. Once Renault speaks, it becomes *common knowledge* that someone has a dirty face. (Assume that it is common knowledge that Renault speaks the truth, at least when he is shocked.) Now let's trace back the consequences.

Consider again the world where only Moe has a dirty face. In that world, once Renault speaks, Moe slaps himself. That's because he learns that he has a dirty face by putting together the clean faces he can see with the fact that someone has a dirty face. (I've been assuming here that it is common knowledge that only Larry, Curly and Moe were in the room to start with. Hopefully that hasn't been too distracting, but it is crucial here.)

Now as a matter of fact, Moe does not immediately slap himself. That suffices to teach everyone something. In particular, it teaches them they were not in the world where only Moe has a dirty face. Of course, they each already knew that, but it is now clear to everyone that they all know it.

Consider next the world where only Curly and Moe have dirty faces. From Curly's perspective in that world, there are two possibilities. Either he and Moe have dirty faces, or only Moe has a dirty face. But we just ruled out that only Moe has a dirty face. So if we were in the world where only Curly and Moe have a dirty face, then Curly should slap himself.

But Curly doesn't slap himself yet. (I'll leave the question of precisely why he doesn't as an exercise; it should be clear given what we've said so far.) So that rules out the possibility that we're in the world where only Curly and Moe have dirty faces. But Larry knew to start with that we were either in the world where all of them have dirty faces, or in the world where only Curly and Moe have dirty faces. So they must be in the world where they all have dirty faces.

At this stage Larry realises this, and slaps himself in the face. At roughly the same time, Curly and Moe also slap themselves in the face. And it's all because of the difference between mutual knowledge and common knowledge.

7.5 Strong and Weak Dominance

The assumption of common knowledge of rationality is a really strong assumption though. The following game makes this very clear.

> **Game 10**
>
> Everyone in a large group selects an integer between 1 and 100 inclusive. The winner of the game is the person whose number is cloest to $^2/_3$ of the average of all of the numbers selected. That is, the payoff for the player who selects closest to $^2/_3$ of the average is 1. (If there is a tie between n players, their payoff is $^1/_n$.) The payoff for everyone

else is 0.

This game can be played with any number of players, but we'll keep things simple by assuming there are just 10. This still gives us too big a game table. We need 10 dimensions, and 100^{10} cells. The latter is not too demanding; but a 10-dimensional representation is tricky on paper. So we'll just describe states of the game.

The first thing to note about the game is that a particular player, let's call her P, can't win if she selects a number between 68 and 100. That's because those numbers can't be 2/3 of the average unless the average is greater than 100. And, of course, the average can't be greater than 100. So those choices are dominated for P.

But we have to be rather careful here. What choice dominates picking, say, 70? We might say that 60 dominates it. After all, 60 could be the best possible play, while 70 could not. But in most circumstances, 60 and 70 will have the same payoff, namely 0. Unless the average is close to 90, or no one else picks around 60, P's payoff will be 0 whether she picks 60 or 70. And the same goes for any alternative to picking 70.

This is all to say that no alternative pick **strongly** dominates picking 70. But several picks do **weakly** dominate it. For instance, picking 64 does. Note that picking 70 can never do better than picking 64, because even if everyone else picks 100, if one player picks 64, the average will be 96.4, so 64 will be closest to 2/3 of the average. So any circumstance where 70 will be a winning play must be one where everyone else picks more than 70. But in those circumstances, picking 64 will win as well. Conversely, picking 64 could do better than picking 70. If everyone else picks 65, picking 64 will win, and picking 70 will lose. So 64 weakly dominates 70. And as we can see, all that really mattered for that argument was that 70 was always going to be higher than 2/3 of the average, so it would be weakly dominated by some numbers that could be closer to 2/3 of the average.

Again, let's list the lessons as we learn them.

- Any selection above 67 is weakly dominated.
- Since rational players do not play weakly dominated strategies, it is irrational to pick any number above 67.

We will, much later on, come back to the assumption that playing weakly dominated strategies is irrational. I think it is true, though it deserves a more careful treatment than we'll give it here. Let's just assume for now it is true.

Now we showed a way that P can win while playing 60. But it has to be said, that it isn't a particularly likely way. It requires the average of the selections to be nearly 90. And that requires a lot of other people to pick high numbers. That is, it requires other people to pick weakly dominated strategies. And that's not very plausible, assuming those other people are rational.

Let's assume, then, that *P* knows that the other players are rational, and hence will not choose weakly dominated strategies. So no other player will choose a number greater than 67. Then the average of what everyone picks can't be greater than 67. So 2/3 of the average can't be greater than 45. So once we remove the weakly dominated strategies, any selection greater than 45 can't be optimal (i.e., it must be considerably greater than 2/3 of the average), and we can give an argument similar to the above argument that it is weakly dominated.

As in the ice-cream game, the trick here is to delete dominated strategies. Once you do that, it is as if you are playing a different game. And in that game, more strategies are in turn dominated. That's because they are strategies that only made sense to play on the assumption that other people played dominated strategies. And, really, it isn't very plausible to assume that people will play dominated strategies. So we should delete the dominated strategies in this new game as well.

And once we do that, we'll find yet more strategies become dominated. Let's say we delete the strategies between 46 and 67. Now the most the average can be is 45. So 2/3 of the average can't be more than 30. So any pick greater than 30 can't be optimal, and so is weakly dominated, so should be deleted. But once those picks are deleted, the average can't be greater than 30, so 2/3 of the average can't be greater than 20, so any pick greater than 20 can't be optimal, and is weakly dominated, so should be deleted. And so on, until every pick greater than 1 is deleted. That's the next lesson from the game.

- The only strategy that survives the iterated deletion of weakly dominated strategies is to select 1.

So it seems rational players, who are playing the game with other rational players, should choose 1 right?

Not so fast! Here's a little tip for anyone playing this game in a large enough group. If you pick 1 you will lose with a probability more or less equal to 1. Just what number will win is harder to predict without knowing more about the group's features, but it won't be 1. Why not? Is it because there are irrational players in any group?

Not necessarily. What's really going on is that the assumptions needed to get to 1 are incredibly strong. Let's go through the argument for getting to 1 in some more detail.

- At the start, 2/3 of the average is at most 67.

- If everyone knows that, and is rational, 2/3 of the average is at most 45.
- If everyone knows that, and is rational, 2/3 of the average is at most 30.
- If everyone knows that, and is rational, 2/3 of the average is at most 20.
- If everyone knows that, and is rational, 2/3 of the average is at most 13.
- If everyone knows that, and is rational, 2/3 of the average is at most 9.
- If everyone knows that, and is rational, 2/3 of the average is at most 6.
- If everyone knows that, and is rational, 2/3 of the average is at most 4.
- If everyone knows that, and is rational, 2/3 of the average is at most 3.
- If everyone knows that, and is rational, 2/3 of the average is at most 2.
- If everyone knows that, and is rational, 2/3 of the average is at most 1.

Note that at every stage we need to make one more assumption about what the players know. By the end we've assumed that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone is rational. (There are 10 iterations of 'everyone knows that' in that sentence, in case you'd rather not count.) And that's really not a plausible assumption.

To put this in perspective, imagine a variant of the Slapville story where there aren't just 3 people, Larry, Curly and Moe, but 10 people. (Perhaps we add the 7 dwarves to the 3 stooges.) And they all have dirty faces. And none of them can see their own face, but Inspector Renault says that at least one of them has a dirty face. It is hard to imagine that this will make any difference at all. For one thing, the computations required to process the new common knowledge are horrendously difficult. (In a way, this is a denial that the players are *perfectly* rational, but it also brings out what a strong assumption that is already.) To assume that they can be done, and that everyone knows they can be done, and everyone knows that everyone knows they can be done, and so on for 8 more steps, is absurd.

7.6 Puzzles about Weak Domination

There is something very odd about choosing strategies that are strongly dominated. And from that, we can naturally deduce that we should delete strategies that are strongly dominated. The iterative deletion of strongly dominated strategies requires not just rationality, but common belief in rationality through as many iterations as there are steps of deletion. Sometimes that will be an implausible assumption, but it often seems reasonable enough in practice.

Weak domination, however, generates principles that are both more puzzling and harder to justify. Some people argue that weakly dominated strategies are perfectly rational to play in games like this one, which we might think of as

Weak Prisoners' Dilemma.

Game 11	l	r
U	1, 1	100, 0
D	0, 100	100, 100

D is weakly dominated by U. But if R is confident that C will play r, then it may be rational to play D. And there's good reason to think that C will play r; the bottom-right corner is a good place for them to both end up. So you might think it is rational to play a weakly dominated strategy.

If we start iteratively deleting weakly dominated strategies, we get even less plausible outcomes.

Game 12	l	m	r
T	2, 2	100, 0	0, 90
M	0, 100	100, 100	100, 95
B	0, 95	95, 100	95, 95

Since B and r are weakly dominated by M and m respectively, we can delete them. And now we're back in a small variant of Game 11, where deleting weakly dominated strategies leads to the $\langle T, l\rangle$ equilibrium. But here it is even more attractive to play $\langle M, m\rangle$. For one thing, it is an equilibrium in an important sense that we'll talk more about later. The important sense is that given what the other player is doing, neither player can do better by changing. So if each player thinks the other player will play their half of $\langle M, m\rangle$, it makes sense for each player to play their half of $\langle M, m\rangle$. All that's true of the $\langle D, R\rangle$ equilibrium in Game 11. But in Game 11, the players couldn't do worse by changing their strategies if they are wrong about what the other players will play. Here they might. We'll have much more to say about this later, but there is an even stronger argument that $\langle M, m\rangle$ is a rational pair of plays in this game than there was that $\langle D, R\rangle$ was a rational pair of plays in Game 11. For roughly these reasons, Robert Stalnaker argues that there is a good sense of rationality (what he calls **perfect rationality**) which requires playing $\langle U, L\rangle$ in Game 11, but which is compatible with playing $\langle M, m\rangle$ in Game 12.

There are other puzzles with iteratively deleting weakly dominated strategies. Surprisingly, it can turn out that the order in which we delete strategies makes a big difference. This point is made in Elon Kohlberg and Jean-Francois Mertens's 1986 *Econometrica* paper "On the Strategic Stability of Equilibria". Here is (a minor variation on) the example they use.

Game 13	l	r
T	2, 2	2, 2
M	1, 0	0, 1
B	0, 1	1, 0

Since B is dominated, it won't be chosen. But once we eliminate B, then for C, r (weakly) dominates l, so only r survives iterative deletion of weakly dominated strategies.

But wait! We could also reason as follows. Since M is dominated, it won't be chosen. But once we eliminate M, then for C, l (weakly) dominates r, so only l survives iterative deletion of weakly dominated strategies. What is going on?

Kohlberg and Mertens suggest that we should focus on strategies that survive *some* process of iterative deletion. Since for player 2, there is an iterative deletion path that l survives, and an iterative deletion path that r survives, then both strategies really survive iterative deletion.

You might be tempted by an alternative take on this example. Perhaps it was wrong to either delete M or to delete B. Perhaps we should say that when we are deleting strategies, the right thing to do is to delete *all* strategies that are dominated at a stroke. So we should simultaneously delete M and B, and then it will be clear that both L and R survive. This won't avoid the problem though, as we can see by a simple three player game.

> **Game 14**
>
> There are three players, 1, 2 and 3. They can each choose one of two options, which we'll label A and B. For player 1 and 2, the payoff structure is easy, they get 2 if they pick A, and 1 if they pick B. For player 3, it is a little more complicated. Player 3 gets:
>
> - 2 if both players 1 and 2 pick A
> - 0 if both players 1 and 2 pick B
> - 1 if players 1 and 2 make opposite picks, and player 3 picks the same thing as player 1.
> - 0 if players 1 and 2 make opposite picks, and player 3 picks the same thing as player 2.

One way we could analyse this is by saying that since B is dominated for both players 1 and 2, they won't pick it. And since player 3's choice doesn't matter if both player 1 and 2 pick A, then it doesn't matter what player 3 picks. But there are other ways we could go as well.

Since B is (strongly) dominated for player 1, we can rule it out. Now player 3 faces the following choice, assuming player 1 picks A. (We'll write player 3's choices on the rows, and player 2's on the columns, and put player 3's payoff first.)

	a	*b*
A	2, 2	1, 1
B	2, 2	0, 1

Now *A* weakly dominates *B*, so it is uniquely rational, we might think, for player 3 to pick *A*.

But wait! Since *b* is (strongly) dominated for player 2, we can rule it out. Now player 3 faces the following choice, assuming player 2 picks *A*. (We'll write player 3's choices on the rows, and player 1's on the columns, and put player 3's payoff first.)

	a	*b*
A	2, 2	0, 1
B	2, 2	1, 1

Now *B* weakly dominates *A*, so it is uniquely rational, we might think, for player 3 to pick *B*.

Now it looks like even we delete every dominated strategy that a player has when we get to that player in the analysis, the order in which we do the deletions still matters. Note though that none of the analysis we've just done seems to undermine the intuitive verdict that player 3 could rationally choose either *A* or *B*. She is going to get 2 whatever she chooses, since the other players will both choose *A*. So this doesn't undermine Kohlberg and Mertens's conclusion that if there is some path of strategy deletion that leads to a strategy being available and no more deletions being possible, then it is (for these purposes) rationally acceptable to choose that strategy.

Chapter 8

Games and Time

So far we've looked just at games where each player makes just one move, and they make it simultaneously. That might not feel like most games that you know about. It isn't how we play, for instance, chess. Instead, most games involve players making multiple moves, and these moves taking place in time. Here is one simple such game. It's not very interesting; you won't have fun games nights playing it. But it is useful to study.

> **Game 15**
>
> There are two players, who we'll call A and B. First A moves, then B, then finally A moves again. Each move involves announcing a number, 1 or 2. A wins if after the three moves, the numbers announced sum to 5. B wins otherwise.

This is a simple zero-sum game. The payoff is either 1 to A and 0 to B, or 1 to B and 0 to A. For simplicity, we'll describe this as A winning or B winning. We'll soon be interested in games with draws, which are a payoff of 1/2 to each player. But for now we're looking at games that someone wins.

Before we go on, it's worth thinking about how you would play this game from each player's perspective. The formal approach we'll eventually take is pretty similar, I think, to the way one thinks about the game.

8.1 Normal Form and Extensive Form

Figure 8.1 is the **extensive form** representation of Game 15. We will use $\mathcal{W}$ to denote that A wins, and $\mathcal{L}$ to denote that B wins.

To read what's going on here, start at the node that isn't filled in, which in this case is the node at the bottom of the tree. The first move is made by A.

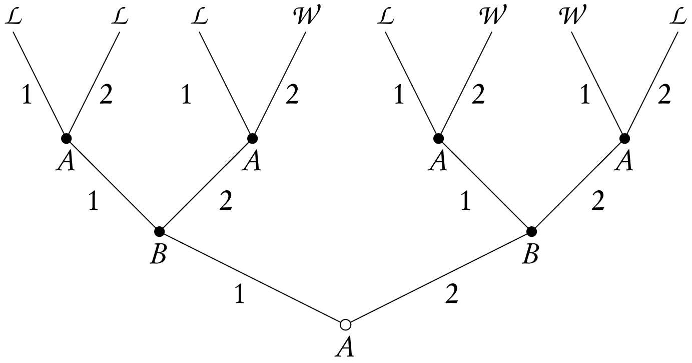


Figure 8.1: Game 15

She chooses to play either 1 or 2. If she plays 1, we go to the left of the chart, and now *B* has a choice. She plays either 1 or 2, and now again *A* has a choice. As it turns out, *A* has the same choice to make whatever *B* does, though there is nothing essential to this. This is just a two-player game, though there is also nothing essential to this. We could easily have let it be the case that a third player moved at this point. Or we could have made it that who moved at this point depended on which move *B* had made. The extensive form representation allows for a lot of flexibility in this respect.

At the top of the diagram is a record of who won. In the more general form, we would put the payoffs to each player here. But when we have a simple zero-sum game, it is easier to just record the payoffs to one player. You should check that the description of who wins and loses in each situation is right. In each case, *A* wins if 2 is played twice, and 1 once.

The extensive form representation form is very convenient for turn taking games. But you might think that it is much less convenient for games where players move simultaneously, as in Prisoners' Dilemma. But it turns out that we can represent that as well, through a small notational innovation. Figure 8.2 is the game tree for Prisoners' Dilemma.

The crucial thing here is the dashed line between the two nodes where P2 (short for Player 2) moves. What this means is that P2 doesn't know which of these nodes she is at when she moves. We normally assume that a player knows what moves are available to her. So we normally only put this kind of dashed line in when it connects two nodes at which the same player moves, and the available moves are the same. When we put in notation like this, we say that the nodes that

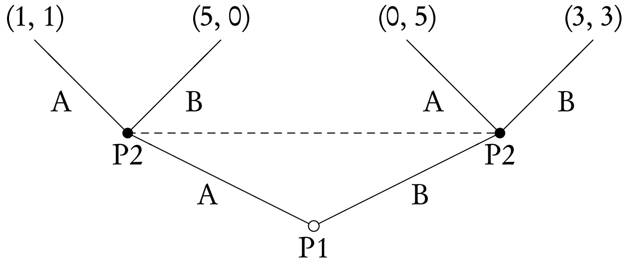


Figure 8.2: Prisoners' Dilemma

are connected form an **information set**. If we don't mark anything, we assume that a node is in a degenerate information set, one that only contains itself.

Strictly speaking, you could regard this tree as a representation of a game where Player 1 goes first, then Player 2 moves, but Player 2 does not know Player 1's move when she moves. But that would be reading, I think, too much into the symbolic representation. What's crucial is that we can represent simultaneous move games in extensive form.

These representations using graphs are knows as **extensive form** representations of games. The represenations using tables that we've used previously are known as **normal form** or **strategic form** representations. The idea is that any game really can be represented as game of one simultaneous move. The 'moves' the players make are the selections of **strategies**. A strategy in this sense is a plan for what to do in any circumstance whatsoever.

Let's do this for the Game 15. A strategy for *A* has three variables. It must specify what she does at the first move, what she does at the second move if *B* plays 1, and what she does at the second move if *B* plays 2. (We're assuming here that *A* will play what she decides to play at the first move. It's possible to drop that assumption, but it results in much more complexity.) So we'll describe *A*'s strategy as $\alpha\beta\gamma$, where α is her move to begin with, β is what she does if *B* plays 1, and γ is what she does if *B* plays 2. A strategy for *B* needs to only have two variables: what to do if *A* plays 1, and what to do if *A* plays 2. So we'll notate her strategy as $\delta\epsilon$, where δ is what she does if *A* plays 1, and ϵ is what she does if *A* plays 2. So *A* has 8 possible strategies, and *B* has 4 possible strategies. Let's record the giant table listing the outcomes if thye play each of those strategies.

Game 15	11	12	21	22
111	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$
112	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{W}$
121	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$
122	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{W}$
211	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{L}$	$\mathcal{W}$
212	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$
221	$\mathcal{W}$	$\mathcal{W}$	$\mathcal{W}$	$\mathcal{W}$
222	$\mathcal{W}$	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{L}$

There is something quite dramatic about this representation. We can see what *A* should play. If her strategy is 221, then whatever strategy *B* plays, *A* wins. So she should play that; it is a (weakly) dominant strategy. This isn't completely obvious from the extended form graph.

Here's a related fact. Note that there are only 8 outcomes of the extended form game, but 32 cells in the table. Each outcome on the tree is represented by multiple cells of the table. Let's say we changed the game so that it finishes in a draw, represented by $\mathcal{D}$, if the numbers picked sum to 3. That just requires changing one thing on the graph; the $\mathcal{L}$ in the top-left corner has to be changed to a $\mathcal{D}$. But it requires making many changes to the table.

Game 15′	11	12	21	22
111	$\mathcal{D}$	$\mathcal{D}$	$\mathcal{L}$	$\mathcal{L}$
112	$\mathcal{D}$	$\mathcal{D}$	$\mathcal{W}$	$\mathcal{W}$
121	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$
122	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{W}$
211	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{L}$	$\mathcal{W}$
212	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$
221	$\mathcal{W}$	$\mathcal{W}$	$\mathcal{W}$	$\mathcal{W}$
222	$\mathcal{W}$	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{L}$

In part because of this fact, i.e., because every change to the value assignment in the extensive form requires making many changes in the values on the normal form, it isn't a coincidence that there's a row containing nothing but $\mathcal{W}$. The following claim about our game can be proved.

> Assign $\mathcal{W}$ and $\mathcal{L}$ in any way you like to the eight outcomes of the extended form game. Then draw table that is the normal form representation of the game. It will either have a row containing nothing but $\mathcal{W}$, i.e., a winning strategy for *A*, or a column containing nothing but $\mathcal{L}$, i.e., a winning strategy for *B*.

We will prove this in the next section, but first we will look at how to 'solve' games like Game 15.

8.2 Backwards Induction

The way to think through games like Game 15 is by working from top to bottom. *A* moves last. Once we get to the point of the last move, there is no tactical decision making needed. *A* knows what payoff she gets from each move, and she simply will take the highest payoff (assuming rationality.)

So let's assume she does that. Let's assume, that is, that *A* does play her best strategy. Then we know three things about what will happen in the game.

- If *A* plays 1, and *B* plays 2, *A* will follow with 2.
- If *A* plays 2, and *B* plays 1, *A* will follow with 2.
- If *A* plays 2, and *B* plays 2, *A* will follow with 1.

Moreover, once we make this assumption there is, in effect, one fewer step in the game. Once *B* moves, the outcome is determined. So let's redraw the game using that assumption, and just listing payoffs after the second move. This will be Figure 8.3

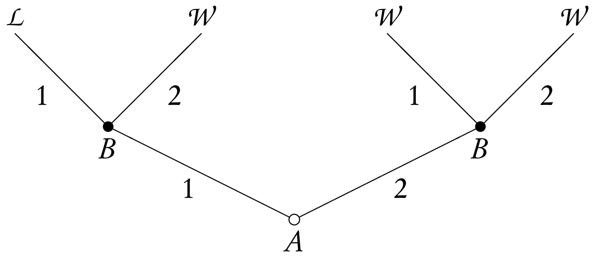


Figure 8.3: Game 15 with last move assumed

Now we can make the same assumption about *B*. Assume that *B* will simply make her best move in this (reduced) game. Since *B* wins if *A* loses, *B*'s best move is to get to $\mathcal{L}$. This assumption then, gives us just one extra constraint.

- If *A* plays 1, *B* will follow with 1.

And, once again, we can replace *B*'s actual movement with a payoff of the game under the assumption that *B* makes the rational move. This gives us an even simpler representation of the game that we see in Figure 8.4.

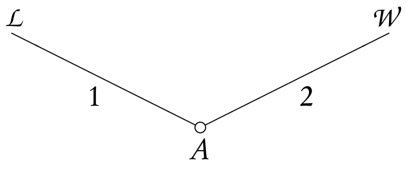


Figure 8.4: Game 15 with last two moves assumed

And from this version of the game, we can draw two more conclusions, assuming A is rational.

- A will play 2 at the first move.
- A will win.

Let's put all of that together. We know A will start with 2, so her strategy will be of the form $2\beta\gamma$. We also know that B doesn't care which strategy she chooses at that point, so we can't make any further reductions. But we do know that if A plays 2 and B plays 1, A will follow with 2. So A's strategy will be of the form 22γ. And we know that if A plays 2 and B plays 2, then A will play 1. So A's strategy will be 221, as we saw on the table.

Note that, as in Game 10, the use of backwards induction here hides a multitude of assumptions. We have to assume each player is rational, and each player knows that, and each player knows that, and so on for at least as many iterations as there are steps in the game. If we didn't have those assumptions, it wouldn't be right to simply replace a huge tree with a single outcome. Moreover, we have to make those assumptions be very modally robust.

We can see this with a slight variant of the game. Let's say this time that the left two outcomes are $\mathcal{D}$. So the graph looks like Figure 8.5.

Now we assume that A makes the optimal move at the last step, so we can replace the top row of outcomes with the outcome that would happen if A moves optimally. This gets us Figure 8.6.

Now assume that A plays 1 on the first round, then B has to move. From 8.6 it looks like B has an easy choice to make. If she plays 1, she gets a draw, if she plays 2, then A wins, i.e., she loses. Since drawing is better than losing, she should play 1 and take the draw.

But why think that playing 2 will lead to A winning? The argument that it did depending on assuming A is perfectly rational. And assuming B is in a position to make this choice, that seems like an unlikely assumption. After all, if A were perfectly rational, she'd have chosen 2, and given herself a chance to force a win.

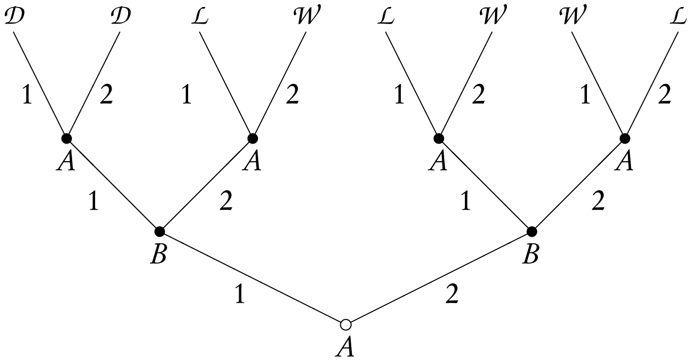


Figure 8.5: Game 15″

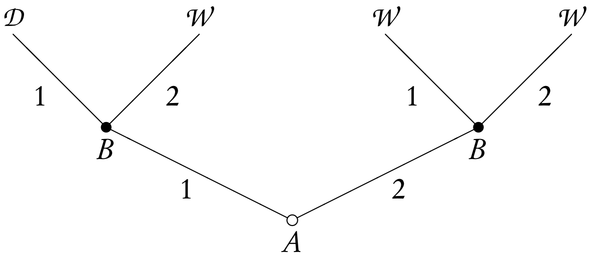


Figure 8.6: Game 15″ with last move assumed

Now you might think that even if *A* isn't perfectly rational, it still is crazy to leave her with an easy winning move. And that's probably a sufficient reason for *B* to accept the draw, i.e., play 1. But the argument that *B* should regard playing 2 as equivalent to choosing defeat seems mistaken. *B* knows that *A* isn't perfectly rational, and she shouldn't assume perfect rationality from here on out.

We will come back to this point, a lot, in subsequent discussions of backwards induction. Note that it is a point that doesn't really arise in the context of normal form games. There we might wonder about whether common knowledge of rationality is a legitimate assumption at the *start* of the game. But once we've settled that, we don't have a further issue to decide about whether it is still a legitimate assumption at later stages of the game.

8.3 Value of Games

Consider games with the following characteristics.

- A and B take turns making moves. We will call each point at which they make a move, or at which the game ends, a **node** of the game.
- At each move, each player knows what moves have been previously made.
- At each move, the players have only finitely many possible moves open.
- The players' preferences over the outcomes are opposite to one another. So if A prefers outcome o_1 to o_2, then B prefers o_2 to o_1, and if A is indifferent between o_1 and o_2, then B is indifferent between them as well.
- A's preferences over the outcomes are complete; for any two outcomes, she either prefers the first, or prefers the second, or is indifferent between them.
- There is a finite limit to the total number of possible moves in the game.

The finiteness assumptions entail that there are only finitely many possible outcomes. So we can order the outcomes by A's preferences. (Strictly speaking, we want to order sets of outcomes that are equivalence classes with respect to the relation that A is indifferent between them.) Assign the outcome A least prefers the value 0, the next outcome the value 1, and so on.

Now we recursively define the **value** of a node as follows.

- The value of a **terminal node**, i.e., a node at which the game ends, is the payoff at that node.
- The value of any node at which A makes a choice is the greatest value of the nodes between which A is choosing to move to.
- The value of any node at which B makes a choice is the least value of the nodes between which B is choosing to move to.

Finally, we say that the value of the game is the value of the initial node, i.e., the node at which the first choice is made. We can prove a number of things about the value of games. The proof will make crucial use of the notion of a **subgame**. In any extensive form game, a **subgame** of a game is the game we get by treating any perfect information node as the initial node of a game, and including the rest of the game 'downstream' from there.

By a perfect information node, I mean a node such that when it is reached, it is common knowledge that it is reached. That's true of all the nodes in most of the games we're looking at, but it isn't true in, for instance, the extensive form version of Prisoners' Dilemma we looked at. Nodes that are in non-degenerate information sets, i.e., information sets that contain other nodes, can't trigger subgames. That's because we typically assume that to play a game, players have to know what game they are playing.

Note that a subgame is really a game, just like a subset is a set. Once we're in the subgame, it doesn't matter a lot how we got there. Indeed, any game we represent is the consequence of some choices by the agent; they are all subgames of the game of life.

The value of a game is the value of one of the terminal nodes

> We prove this by induction on the length of games. (The length of a game is the *maximum* number of moves needed to reach a terminal node. We've only looked so far at games where every path takes the same number of moves to reach a conclusion, but that's not a compulsory feature of games.)
>
> If the game has zero moves, then it has just one node, and its value is the value of that node. And that node is a terminal node, so the value is the value of a terminal node.
>
> Now assume that the claim is true for any game of length k or less, and consider an arbitrary game of length $k+1$ The first node of the game consists of a choice about which path to go down. So the value of the initial node is the value of one of the subsequent nodes. Once that choice is made, we are in a subgame of length k, no matter which choice is made. By the inductive hypothesis, the value of that subgame is the value of one of its terminal nodes. So the value of the game, which is the value of one of the immediate subsequent nodes to the initial node, is the value of one of its terminal nodes.

***A* can guarantee that the outcome of the game is at least the value of the game.**

> Again, we prove this by induction on the length of games. It is trivial for games of length 0. So assume it is true for all games of length at most k, and we'll prove it for games of length $k+1$. The initial node of such a game is either a move by A or a move by B. We will consider these two cases separately.
>
> Assume that A makes the first move. Then the value of the initial node is the maximum value of any immediate successor node. So A can select to go to the node with the same value as the value of the game. Then we're in a subgame of length k. By the inductive assumption, in that game A can guarantee that the outcome is at least the value of the subgame. And since the value of that node is the

> value of the subgame, so it is also the value of the initial node, i.e., the value of the initial game. So by choosing that node, and starting that subgame, *A* can guarantee that the outcome is at least the value of the game.
>
> Now assume that *B* makes the first move. *B* can choose the node with the least value of the available choices. Then, as above, we'll be in a subgame in which (by the inductive hypothesis) *B* can guarantee has an outcome which is at most its value. That is, *B* can guarantee the outcome of the game is at most the value of the initial node of the subgame. And since *B* can guarantee that that subgame is played, *B* can guarantee that the game has an outcome of at most its value.

***B* can guarantee that the outcome of the game is at most the value of the game.**

> The proof of this exactly parallels the previous proof, and the details are left as an exercise.

Let's note a couple of consequences of these theorems.

First, assume that the rationality of each player is common knowledge, and that it is also common knowledge that this will persist throughout the game. Then the kind of backwards induction argument we used is discussing Game 15 will show that the outcome of the game will be the value of the game. That's because if *A* is rational and knows this much about *B*, the outcome won't be lower than the value, and if *B* is rational and knows this much about *A*, the outcome won't be greater than the value.

Second, these theorems have many applications to real-life games. Both chess and checkers, for instance, satisfy the conditions we listed. The only condition that is not immediately obvious in each case is that the game ends in finite time. But the rules for draws in each game guarantee that is true. (Exercise: Prove this!) Since these games end with White win, Black win or draw, the value of the game must be one of those three outcomes.

In the case of checkers, we know what the value of the game is. It is a draw. This was proved by the Chinook project at the University of Alberta. We don't yet know what the value of chess is, but it is probably also a draw. Given how many possible moves there are in chess, and in principle how long games can go on, it is hard to believe that chess will be 'solved' any time soon. But advances in chess computers may be able to justify to everyone's satisfaction a particular solution, even if we can't prove that is the value of chess.

Chapter 9

Best Responses

9.1 Best Responses

In the last two chapters, we just used the concept of a dominating strategy to solve various games. But we can do better by bringing in the primary tool from decision theory, namely the notion of an expected payoff, to rule out various strategies. The core notion we will use is that of a **Best Response**.

Best Response A strategy s_i is a best response for player i iff there is some probability distribution Pr over the possible strategies of other players such that playing s_i maximises i's *expected* payoff, given Pr. (Note that we're using 'maximise' in such a way that it allows that other strategies do just as well; it just rules out other strategies doing better.)

Let's look at an example of this, using a game we have already seen.

Game 5	l	r
U	3, 0	0, 0
M	2, 0	2, 0
D	0, 0	3, 0

We will just look at things from the perspective of R, the player who chooses the row. What we want to show is that *all three* of the possible moves here are best responses.

It is clear that U is a best response. Set $\Pr(l) = 1, \Pr(r) = 0$. Then $E(U) = 3, E(M) = 2, E(D) = 0$. It is also clear that D is a best response. Set $\Pr(l) = 0, \Pr(r) = 1$. Then $E(U) = 0, E(M) = 2, E(D) = 3$.

The striking thing is that M can also be a best response. Set $\Pr(l) = \Pr(r) = 1/2$. Then $E(U) = E(D) = 3/2$. But $E(M) = 2$, which is greater than $3/2$. So if R

thinks it is equally likely that C will play either l or r, then R maximises expected utility by playing M. Of course, she doesn't maximise actual utility. Maximising actual utility requires making a gamble on which choice C will make. That isn't always wise; it might be best to take the safe option.

It's very plausible that agents should play best responses. If a move is not a best response, then there is no way that it maximises expected utility, no matter what one's views of the other players are. And one should maximise expected utility.

Making this restriction on what agents do rules out some strategies that are permitted by any kind of dominance reasoning, whether we use strong or weak dominance, and whether or not we iterate. We can see this with the following game.

Game 16	l	r
U	3, 3	0, 0
M	1, 1	1, 1
D	0, 0	3, 3

Consider things from R's perspective. The probability that C will play l is x, for some x, and the probability that she will play r is $1-x$. So $E(U)=3x, E(D)=3(1-x)$ and $E(M)=1$. If M is to be a best response, then we must have $E(M) \geq E(U)$, and $E(M) \geq E(D)$. The first condition entails that $1 \geq 3x$, i.e., $x \leq {}^{1}/_{3}$. The second condition entails that $1 \geq 3(1-x)$, i.e., $x \geq {}^{2}/_{3}$. But these two conditions can't both be satisfied, so M is not a best response under any circumstances.

But nor is M even weakly dominated. Since M sometimes does better than U, and sometimes does better than D, it is not dominated by either. Moreover, neither of C's strategies is weakly dominated by the other. So eliminating weakly dominated strategies removes no strategies whatsoever from the game. Hence M is NWDAI, and hence NSDAI, NWD and NSD, but it is not BR.

9.2 Iterated Best Responses

Some best responses are pretty crazy when playing against a rational opponent. Consider the following game from R's perspective.

Game 17	L	r
U	5, 5	0, -5
D	0, 5	2, -5

In this game, D is a best response. It does best if C chooses r. But why on earth would C do that? C gets 5 for sure if she chooses L, and -5 for sure if she chooses r. Assuming the weakest possible rationality constraint on C, she won't choose a sure loss over a sure gain. So given that, C should choose U.

Of course, we could have shown that with just considering domination. Note that U is both NSDAI and NWDAI, while D has neither of these properties. The following example is a little more complex.

Game 18	l	m	r
U	1, 3	0, 1	1, 0
D	0, 0	1, 1	0, 3

In this game neither player has a dominated move. So just using domination techniques can't get us closer to solving the game. And, from R's perspective, thinking about best responses doesn't help either. U is a best response if C is going to play l or r with probability at least 1/2, and D is best response if C is going to play m with probability at least 1/2.

But note that m is not a best response for C. The argument for this is just the argument we used in Game 16. Now let's assume, when thinking from R's perspective, that C will play a best response. That is, we're assuming C will play either l or r. Given that, the best thing for R to do is to play U.

More carefully, if R knows that C is rational, and if R knows that rational agents always play best responses, then R has a compelling reason to play U. This suggests an interesting status that some strategies may have.

BRBR That is, is a **B**est **R**esponse to a **B**est **R**esponse.

We can go on further. Think about Game 18 from C's perspective. Assuming R is rational, and knows C is rational, then R has a compelling reason to play U. And if R plays U, the best response for C is l. This doesn't mean that l is the only best response; r is also a best response. Nor does it mean l is the only best response to a best response; r is the best response to D, which as we showed above is a best response. What's really true is that l is the only best response to a best response to a best response.

That means that if C knows that R knows that C is rational, then C has a strong reason to play l. We could designate this with a new status, perhaps BRBRBR. But at this point it's best to simply iterate to infinity.

BRBRI That is, is a **B**est **R**esponse to a **B**est **R**esponse to a Best Response, and so on to **I**nfinity.

We say that p is **mutual knowledge** if every player in the game knows it. We say that p is **common knowledge** if everyone knows it, and everyone knows everyone knows it, and so on. It seems plausible that in any game where it is mutual knowledge that everyone is rational, agents should only play BRBR strategies. And at least there's a case to be made that in a game where it is common knowledge that the players are rational, players should play BRBRI strategies. (Though we will come back to this point repeatedly, especially in the context of extensive form games.)

In a two-player game, for a strategy s_0 to be BRBRI, it must be the best response to some strategy s_1, which is the best response to some strategy s_2, which is the best response to some strategy s_3, etc. But we are assuming that there are only finitely many strategy choices available. So how can we get such an infinite chain going?

The answer, of course, is to repeat ourselves. As long as we get some kind of loop, we can extend such a chain forever, by keeping on circling around the loop. And the simplest loop will have just two steps in it. So consider any pair of strategies $\langle s_0, s_1 \rangle$ such that s_0 is the best response to s_1, and s_1 is the best response to s_0. In that case, each strategy will be BRBRI, since we can run around that two-step 'loop' forever, with each stop on the loop being a strategy which is a best response to the strategy we previous stopped at.

When a pair of strategies fit together nicely like this, we say they are a **Nash Equilibrium**. More generally, in an n-player game, we use the following definition.

NE Some strategies $s_1, ..., s_n$ in an n-player game form a **N**ash **E**quilibrium iff for each i, s_i is a best response to the strategies played by the other $n - 1$ players. That is, iff player i cannot do better by playing any alternative strategy to s_i, given that the other players are playing what they actually do.

The 'Nash' in Nash Equilibrium is in honour of John Nash, who developed the concept, and proved some striking mathematical results concerning it. (You may remember that Nash was the subject of a bio-pic some years back, 'A Beautiful Mind'. The movie required believing that Russell Crowe was a mad genius, and didn't properly deploy the notion of Nash Equilibrium, so maybe it is best if we keep our contact with Nash to the textbooks.)

We can prove that being NE is a stricter requirement than being BRBRI with the following game.

Game 19	l	r
U	1, 0	0, 1
D	0, 1	1, 0

None of the (pure) strategies U, D, l or r are NE. That's because there's no NE pair we can make out of those four. And that's fairly obvious from the fact that whichever corner of the table we end up in, one of the players would have done better by swapping their strategy.

But note that each of the four strategies is BRBRI. U is a best response to l, which is a best response to D, which is a best response to r, which is a best response to U, which is ....

The point here should be clear once we think about how we got from the idea of BRBRI to the idea of NE. We wanted a 'loop' of strategies, such that each was a best response to the strategy before it. NE was what we got when we had a loop of length 2. But there are loops which are longer than that; for example, there are loops of length 4. And any loop is sufficient for the strategies on the loop to be BRBRI. And these strategies need not be NE.

9.3 Nash Equilibrium in Simultaneous Move Games

Let's return to Game 19. It looks at first like there won't be any Nash Equilibrium strategies in this game. That would be unfortunate; all of our statuses so far are exemplified by at least one strategy in each game.

But that would be too quick. It leaves out a large class of strategies. Game theorists say that as well as simply choosing to play U, or choosing to play D, R has another choice. She can play a **mixed strategy**. A mixed strategy is where the player plays different pure strategies with different probabilities. We'll come back very soon to what we might possibly *mean* by 'probability' here, but for now let's explore the consequences for the existence of Nash Equilibria.

Let's assume that R plays U with probability $1/2$, and D with probability $1/2$. And similarly, assume that C plays l with probability $1/2$, and r with probability $1/2$. Can either player do better by deviating from these strategies?

Let's look at it first from C's perspective. If she plays l, then her expected return is given by the following equation.

$$
\begin{aligned}
E(L) = &\text{ Prob that } R \text{ plays } U \times \text{Payoff of } \langle U, l\rangle \\
&+ \text{Prob that } R \text{ plays } D \times \text{Payoff of } \langle D, l\rangle \\
= &\ 1/2 \times 0 + 1/2 \times 1 \\
= &\ 1/2
\end{aligned}
$$

And the expected return of playing r is given by the following equation.

$$\begin{aligned}E(R) &= \text{Prob that } R \text{ plays } U \times \text{Payoff of } \langle U, r\rangle \\ &\quad + \text{Prob that } R \text{ plays } D \times \text{Payoff of } \langle D, r\rangle \\ &= 1/2 \times 1 + 1/2 \times 0 \\ &= 1/2\end{aligned}$$

Let M_x be the mixed strategy of playing L with probability x, and R with probability $1 - x$. Then the expected value of M_x is given by the following equation.

$$\begin{aligned}E(M_x) &= \Pr(l)E(l) + \Pr(r)E(r) \\ &= x/2 + 1-x/2 \\ &= 1/2\end{aligned}$$

So whichever strategy C adopts, whether it is l, r or one of the continuum many values of M_x, she'll have an expected payout of 1/2. That means that she can't do any better by playing any alternative to $M_{1/2}$. Of course, that's for the rather boring reason that any strategy is as good as any other at this point.

When we're discussing R's strategies, we'll say that M_x is the strategy of playing U with probability x, and D with probability $1 - x$. A similar argument shows that given that C is playing $M_{1/2}$, all strategies are as good as each other for R. That means that the pair $\langle M_{1/2}, M_{1/2}\rangle$ is a Nash Equilibrium. Each player does as well as they can playing $M_{1/2}$ given that the other player is playing $M_{1/2}$. And that's the definition of a Nash Equilibrium.

It turns out that for any game with a finite number of choices for each player, there is always at least one Nash Equilibrium, if we include mixed strategies. The proof of this is beyond the scope of these notes, however.

Rather than using ad hoc naming conventions like M_x, it would be good to have better ways of referring to mixed strategies. I'll use the following (fairly standard) notation. If a player's choices for pure strategies are $s_1, s_2, ..., s_n$, then the vector $\langle x_1, x_2, ..., x_n\rangle$ will represent the mixed strategy of playing s_i with probability x_i. If the game is represented on a table, we'll let the first (i.e., leftmost) column be C's strategy s_1, the second column be her strategy s_2, and so on. And we'll let the first (i.e., highest) row be R's strategy s_1, the second row be her strategy s_2, and so on. This notation will need to get more complicated when we consider games in extensive form, but in fact we usually use mixed strategies for games displayed in strategic form, so this isn't a huge loss.

9.4 Mixtures and Dominance

Now that we have the notion of a mixed strategy, we need to revise a little bit about what we said about dominant strategies. Recall that we used the Game 16 to show that some strategies which were not dominated were nevertheless not best responses.

Game 16	l	r
U	3, 3	0, 0
M	1, 1	1, 1
D	0, 0	3, 3

Now compare the strategy M to the strategy $S = \langle 1/2, 0, 1/2 \rangle$. If C plays l, then M returns 1, but the mixed strategy has an expected return of 1.5. If C plays r, then M returns 1, but S has an expected return of 1.5. Indeed, if C plays any mixture of l and r, then M returns 1, but S still has an expected return of 1.5. In effect, M is a dominated strategy; it is dominated by S.

So when we are talking about whether a strategy is dominated, we have to distinguish the scope of the tacit quantifier. Remember, 'dominated' means 'dominated by something'. If we are only quantifying over pure strategies, then M is not dominated. If we are quantifying over mixed strategies as well, then M is dominated. At least some authors use 'dominated' with this more expansive quantifier domain in mind, so M is dominated because S dominates it. There is a nice advantage to doing this. (There are also costs; it is harder to tell just by looking whether strategies are dominated in this more expansive sense.) It means we can distinguish nicely between what I've called a Best Strategy, which is really a strategy that is not strongly dominated, and what we might call Perfect Strategies, which are strategies that are not weakly dominated in this sense. We'll come back to this when we look at Stalnaker's work in a few chapters.

9.5 What is a Mixed Strategy?

So far we've noted that there are some games that don't have pure strategy Nash Equilibria. And we said that if we expand the space of available strategies to include mixed strategies, then those games do have Nash Equilibria. In fact we've said, though not gone close to proving this, that all finite games have at least one Nash Equilibrium solution, once we allow that agents can play mixed strategies.

But what does it mean to 'play a mixed strategy'? As game theorists sometimes put it, how should be **interpret** talk of mixed strategies. It turns out the options here are very similar to the candidate 'interpretations' of probability. (See the SEP entry on interpretations of probability for more on this debate, if you're interested.) Basically the interpretations can be classified as either **metaphysical** or **epistemological**. We're going to start with one of the metaphysical interpretations, then look at a few epistemological interpretations, and finally return to some more mundane metaphysical interpretations.

The most obvious, and natural, interpretation uses objective chances. What it means to play the strategy $\langle x, 1-x \rangle$ is to grab some chance device that goes into one state with chance x, and another state with chance $1-x$, see which state it goes into, then play the s_1 if it is in the first state, and s_2 if it is in the second state. Consider, for instance, the game Rock, Paper, Scissors, here represented as **Game 20**.

Game 20	Rock	Paper	Scissors
Rock	0, 0	-1, 1	1, -1
Paper	1, -1	0, 0	-1, 1
Scissors	-1, 1	1, -1	0, 0

For each player, the equilibrium strategy is to play $\langle 1/3, 1/3, 1/3 \rangle$. (Exercise: Verify this!) The chance interpretation of this mixed strategy is that the player takes some randomising device, say a die, that has a 1/3 chance of coming up in one of three states. Perhaps the player rolls the die and plays Rock if it lands 1 or 2, Paper if it lands 3 or 4, Scissors if it lands 5 or 6.

A slightly more elegant version of this strategy involves the game Matching Pennies. We've seen the formal version of this game before, but the informal version is fun as well. Basically each player reveals a penny, and Row wins if they are both heads up or both tails up, and Column wins if one coin is heads up and the other tails up. Apparently this was a source of some schoolyard amusement before students had things like Angry Birds, or football, to play. As I said, we've seen the game table before, though not with these labels.

Game 21	Heads	Tails
Heads	1, -1	-1, 1
Tails	-1, 1	1, -1

Again, the only equilibrium solution is for each player to play $\langle 1/2, 1/2 \rangle$. And here the chance interpretation of this strategy is that each player plays by simply flipping their coin, and letting it land where it may.

But obviously it is very hard to procure a chance device on the spot for any mixed strategy one might choose to play. How should we interpret a mixed strategy then? The natural move is to opt for some kind of *epistemological* interpretation of mixtures.

One option is a straightforward subjectivist interpretation of the relevant probabilities. So Row plays a mixed strategy $\langle x, 1-x\rangle$ iff Column's subjective probability that Row is playing s_1 with is x, and her subjective probability that Row is playing s_2 is $1-x$. One does hear game theorists offer such subjective interpretations of mixed strategies, but actually they don't seem to make a lot of sense. For one thing, it's hard to say how Column's credences should be in any sense a *strategy* for Row, unless Row has Jedi mind-control powers. And if Row does have Jedi mind-control powers, then she shouldn't settle for any kind of mixed strategy equilibrium. In Rock, Paper, Scissors, for instance, she should follow the strategy of playing Rock and using her Jedi mind-control powers to get Column to think she's playing Paper.

Perhaps we can retain a subjectivist interpretation if we change who the subject is. Frank Arntzenius, in a recent paper called "No Regrets", offers a different kind of subjectivist interpretation. He says that an agent plays a mixed strategy $\langle x, 1-x\rangle$ if her credences at the end of rational deliberation are that she'll play s_1 with probability x, and s_2 with probability $1-x$. He admits there are oddities with this interpretation. In particular, he notes that it means that if you take some kind of equilibrium theory to be the theory of rationality (as he does), then our theory of rationality turns out to be a theory of what one should believe one will do, not what one will do. This feels a little like changing the subject.

Could some kind of objective version of an epistemological interpretation do any better? Perhaps we could say that to play $\langle x, 1-x\rangle$ is to act in such a way that the objectively rational thing to believe about the agent is that she'll play s_1 with probability x? Arguably, the objective chance interpretation is a version of this; given the Principal Principle, the rational thing to believe about a person who uses a randomising device that comes up s_1 with chance x is that they'll play s_1 with probability x. But beyond that, it is hard to see what advantage the view has. After all, if it is an available strategy to make rational people think that you're playing s_1 with probability x, in Rock, Paper, Scissors you should make them think you're likely playing Paper when you're actually playing Rock. So it's hard to see how the equilibrium solution is rational.

In Ken Binmore's decision theory textbook *Playing for Real*, he seems to endorse something like the objective epistemological interpretation of mixed strategies.

> Suppose that we deny Eve access to a randomizing device when she plays Matching Pennies with Adam. Is she now doomed to lose? Not if she knows her Shakespeare well! She can then make each choice of *head* or *tail* contingent on whether there is an odd or even number of speeches in the successive scenes of *Titus Andronicus*. Of course,

> Adam might in principle guess that this is what she is doing—but how likely is this? He would have to know her initial state of mind with a quite absurd precision in order to setle on such a hypothesis. Indeed, I don't know myself why I chose *Titus Andronicus* from all Shakespeare's plays . . . To outguess me in such a manner, Adam would need to know my own mind better than I know it myself. (Binmore 2006, 185).

But why is the likelihood that Adam will figure out Eve's decision a component of Eve's *strategy*? Either Eve has some control over the probabilities Adam will assign to her making various choices, or she doesn't. If she doesn't, then she doesn't have the power to play any mixed strategy interpreted this way. If she does, she shouldn't use them to give Adam *true* beliefs about her likely decisions. So it's hard to see the advantages.

Perhaps then the detour through epistemological interpretations was a red herring. And this leads us back to two more metaphysical interpretations, both of them involving frequencies.

One of these interpretations is particularly important in biology, and we will return to this idea much later in the course. The idea is that a species plays a mixed strategy $\langle x, 1-x \rangle$ iff x of its population plays s_1, and $1-x$ of its population plays s_2. Once we're looking at population level 'strategies' we've obviously moved a little away from the focus on *rational* game playing that has been the focus of the course to date. It is not obvious that it even makes sense to assign agency to populations. And presumably the way the population implements the policy of having this frequency distribution of strategies is by having some randomising device that sorts individual organisims into one of two types. So perhaps this isn't really an alternative to the objective chance interpretation either.

The other big alternative is hinted at in Binmore's discussion. Note that he refers to 'each choice of *head* or *tail*'. This implicates at least that there will be more than one. So what he's really interested in is the case where Adam and Eve play Matching Pennies repeatedly. In this case, we might say that playing a mixed strategy $\langle x, 1-x \rangle$ is playing s_1 in x of the repeated games, and playing s_2 in $1-x$ of the repeated games. (Actually, we might want something more specific than that, but saying just what we need is hard. See the Stanford Encyclopedia of Philosophy entry on "Chance versus Randomness" for more discussion of the difficulties here.)

But it's odd to bring repeated games in here. It looks suspiciously like changing the subject. What we care about is what we should do in this very game, not in a long sequence of games. Put another way, we should in the first instance be looking for a rule about what to do in a (known to be) one-shot game. What should we do in such a game? Considerations about what would happen were we to repeat the game seem irrelevant to that.

The issues about repeated games are complicated, and it is worth spending some time going over them directly.

9.6 Nash, Best Response and Repeated Games

Here's a hypothesis; we're not worried for now whether it is true, though that's a question we'll want to return to. The hypothesis is that in any game where there is common knowledge of rationality, any strategy which is BRBRI (i.e., that is a best response to a best response to a best response ...) can be rationally played.

Now here's an objection to that hypothesis. Consider repeated games of Rock, Paper, Scissors. In any given game, playing Rock is BRBRI. That's because playing Rock is a best response to playing Scissors, which is a best response to playing Paper, which is a best response to playing Rock, and so on. But playing Rock repeatedly is dumb. If it weren't dumb, the following scene (from The Simpsons episode "The Front" (April 15, 1993)) wouldn't be funny.

> Lisa: Look, there's only one way to settle this. Rock-paper-scissors.
> Lisa's brain: Poor predictable Bart. Always takes 'rock'.
> Bart's brain: Good ol' 'rock'. Nuthin' beats that!
> Bart: Rock!
> Lisa: Paper.
> Bart: D'oh!

Since Bart's strategy is BRBRI, and is irrational, it follows that the hypothesis is false.

I think that objection is too quick, but it takes some time to say why. Let's think a bit about where Bart goes wrong. Imagine he and Lisa play Rock, Paper, Scissors (hereafter RPS) many many times. At first Lisa plays the mixed strategy $\langle 1/3, 1/3, 1/3 \rangle$, and Bart plays Rock. So each of them have an expected return of 0. By the 11þ round, Lisa has figured out what Bart is playing, and plays Paper, for an expected return of 1, while Bart has an expected return of -1.

Now let's think about when Bart goes wrong. To do this, I'm going to assume that Bart is rational. This is clearly false; Bart is obviously not playing rationally. But I want to see just where the assumption of rationality leads to a contradiction. Let's say we knew in advance nothing but that Bart was going to play Rock on the 11th round. Since we're still assuming Bart is rational, it follows that playing Rock is a best response given his credences over Lisa's moves. Without I hope loss of generality, let's assume Lisa is still playing $\langle 1/3, 1/3, 1/3 \rangle$. Then Bart's expected return is 0, like for any other strategy, so it's fine to play Rock.

But of course Bart doesn't just play Rock in the 11th round. He also plays in the previous ten rounds. And that means that Lisa won't still be playing $\langle 1/3, 1/3, 1/3 \rangle$ by the time we get to the 11th round. Instead, she'll be playing Paper. So Bart's expected return in the 11th round is not 0, it is (roughly) -1.

I think something follows from those reflections. When we added in information about Bart's play in the first ten rounds, Bart's expected return in the 11th round dropped. So I conclude that there was a long-term cost to his play in the first 10 rounds. When he played Rock all those times, his expected return *in that round* was 0. But he incurred a long-term cost by doing so. That long-term cost isn't properly represented in the matricies for each round of the game. When we include it, it no longer becomes clear that Rock is BRBRI in each round.

A natural question then is, what really is the payoff table for each round of RPS. The existing table isn't a payoff table for two reasons. First, it lists outcomes, not valuations of outcomes. And second, it only lists one kind of outcome, the short-run winnings, not the long-run consequences of any given strategy. What we really need is a valuation function over long-run consequences.

So what is the payoff table for each round of the game? I think that's just much too hard a question to answer. What we can say with some precision is what the short-run outcome table is for each round. And summing short-run outcomes gives us long-run outcomes, which hopefully will correlate in some sensible way with payoffs. But determining the long-run outcomes of any given strategy is much too hard.

And that's one reason to care about both mixed strategies, and the existence of equilibria. Sometimes the best we can do in specifying a game is to specify the short-run outcomes, and say that the long-run outcomes are sums of them. Now the following hypothesis is clearly false: In any long-run game, it is rationally permissible to, at each stage, play any strategy which would be BRBRI if the short-run outcomes were the total payoffs. The Bart and Lisa example refutes that hypothesis. But it doesn't refute the hypothesis that I started this section with, namely that any BRBRI strategy is rationally acceptable.

9.7 Equilibrium and Ratification

We'll end today with another question about the philosophical significance of equilibrium concepts. A large number of fairly orthodox game theorists typically accept the following two propositions.

- It is strictly better to defect than co-operate in Prisoners' Dilemma, and in general one should always take strictly dominating options.
- In some games, such as Rock-Paper-Scissors, the best thing to do is to play a mixed strategy. In those games, playing a mixed strategy is preferably to

playing any pure strategy.

In general, anyone who thinks there is something normatively significant in playing Nash Equilibrium strategies will have to accept something like those two claims. But if you look at the two biggest positions in philosophical decision theory, they are hard to reconcile. Evidential Decision Theorists may accept the second, but will reject the first. On Evidential Decision Theory, it may be best to accept a dominated option if, as in Newcomb's Problem, the states are evidentially connected to one's choice. And Causal Decision Theorists will accept the first, but not the second. On Causal Decision Theory, the expected value of a mixed strategy is just the (weighted) average value of the strategies being mixed, and the only rule is to maximise expected value, so the mixed strategy can't be preferable to each of the elements of the mixture.

Some of the tension is resolved if we add ratificationist decision theories to our menu of choices. The idea behind ratificationism is that only *ratifiable* decisions are rationally allowed. A decision is ratifiable if it maximises expected utility conditional on that decision being taken. We can add a ratifiability clause to Evidential Decision Theory. That's what Richard Jeffrey originally proposed, as a way of endorsing the idea that we should take both boxes in Newcomb's Problem (or at least in real-life Newcomb-like cases) without giving up most of evidential decision theory. Alternatively, we can add a ratifiability clause to Causal Decision Theory. Frank Arntzenius has recently suggested such a move.

If we add a ratifiability clause to Evidential Decision Theory, we get the result that rational agents should take both boxes. That's because only it is ratifiable. We computed earlier the expected utility of each choice according to Evidential Decision Theory, and concluded that the utility of taking just one box was higher. But now look what happens if we conditionalise on the hypothesis that we'll take just one box. (For simplicity, we'll again assume $1 is worth 1 util.) It is easy enough to see that taking both boxes is better.

$$\begin{aligned}
\text{Pr(Million in opaque box|Take one box)} &= 0.99 \therefore \\
V(\text{Take one box|Take one box}) &= 0.99 \times 1,000,000 + 0.01 \times 0 \\
&= 990,000 \\
V(\text{Take both box|Take one box}) &= 0.99 \times 1,001,000 + 0.01 \times 1,000 \\
&= 991,000
\end{aligned}$$

But there is something very odd about this way of putting things. It requires thinking about the expected value of an action conditional on something that entails the action is not taken. In Newcomb's Problem we can sort of make sense of this; we use the conditional assumption that we're taking one box to seed the probability of the demon doing certain actions, then we run the calculations from there. But I don't see any reason to think that we should, in general, be able to make sense of this notion.

A better approach, I think, is to mix ratificationism with Causal Decision Theory. (This isn't to say it's the right approach; it might be best to skip ratifiability clauses altogether) This lets us solve problems like the following two-step Newcomb problem. In this game the player has a choice between taking $2,000 or playing Newcomb's Problem. If the player does the latter, she must choose one box or two. We will assume the demon is very very accurate; given that the player is choosing φ, the probability that the demon will predict φ is 0.9999. Now let's work through the values of taking each of the options. (We'll use p_i is the proposition that the demon predicts that i boxes will be taken. And we'll use T_i as shorthand for *Take i boxes*. And we'll assume, again that a pound is worth a util.)

$$\begin{aligned}
U(T_1|T_1) &= \Pr(T_1 \mathrel{\Box\!\!\rightarrow} p_1|T_1)U(T_1 \wedge p_1) + \Pr(T_1 \mathrel{\Box\!\!\rightarrow} p_2|T_1)U(T_1 \wedge p_2)\\
&= 0.9999 \times 1,000,000 + 0.0001 \times 0\\
&= 999,900\\
U(T_2|T_1) &= \Pr(T_2 \mathrel{\Box\!\!\rightarrow} p_1|T_1)U(T_2 \wedge p_1) + \Pr(T_2 \mathrel{\Box\!\!\rightarrow} p_2|T_1)U(T_2 \wedge p_2)\\
&= 0.9999 \times 1,001,000 + 0.0001 \times 1,000\\
&= 1,001,900\\
U(T_1|T_2) &= \Pr(T_1 \mathrel{\Box\!\!\rightarrow} p_1|T_2)U(T_1 \wedge p_1) + \Pr(T_1 \mathrel{\Box\!\!\rightarrow} p_2|T_2)U(T_1 \wedge p_2)\\
&= 0.0001 \times 1,000,000 + 0.9999 \times 0\\
&= 100\\
U(T_2|T_2) &= \Pr(T_2 \mathrel{\Box\!\!\rightarrow} p_1|T_2)U(T_2 \wedge p_1) + \Pr(T_2 \mathrel{\Box\!\!\rightarrow} p_2|T_2)U(T_2 \wedge p_2)\\
&= 0.0001 \times 1,001,000 + 0.9999 \times 1,000\\
&= 1,100
\end{aligned}$$

The important thing to note about this calculation is that $\Pr(T_2 \mathrel{\Box\!\!\rightarrow} p_1|T_1)$ is very high, 0.9999 in our version of the game. What this says is that once we've assumed T_1, then the counterfactual $T_2 \mathrel{\Box\!\!\rightarrow} p_1$ is very very probable. That is, given that we're taking 1 box, it is very probable that if we had taken 2 boxes, there would still have been money in the opaque box. But that's just what Newcomb's problem suggests.

Note that neither T_1 nor T_2 is ratifiable. Given T_1, the player would be better with T_2. (The expected value of taking both boxes would be 1,001,900, as compared to an expected value of 999,900 for taking one box.) And given T_2, the

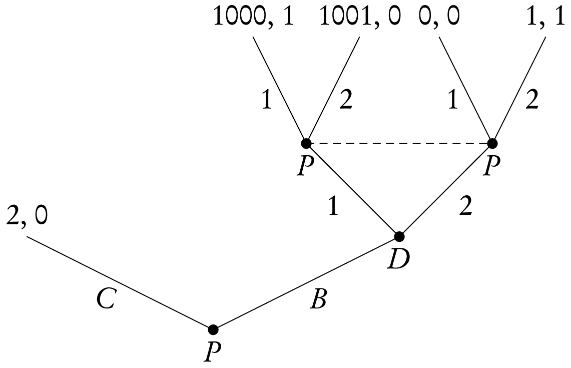


Figure 9.1: Game 22

player would be better with simply taking the $2,000 and not playing, since the expected payout of T_2 is a mere 1,100. But taking $2,000 is ratifiable. Given that the player is doing this, no other option is better. After all, if they are the kind of player who is moved by the reasoning that leads to taking the $2,000, then they are almost certainly two boxers, and so probably the opaque box would be empty. So the only ratifiable decision is to take $2,000. This ratificationist argument is, I think, intuitively plausible.

Note too that it is the very same conclusion that we reach through using a form of backward induction. Let's set up the game as a chart. We'll use *P* for the player, and *D* for the demon. At move 1, *P*'s choices are *C*ash or *B*ox. After that, we'll use 1 and 2 for the obvious meanings; predictions for the demon, boxes for the players. The payoffs are player first, demon second. To make the numbers easier to read, we'll set 1 util as being worth $1,000. And we'll call the extended form of this **Game 22**.

Note crucially the dotted line between the player's choices in the top right. Although the player makes her choice *after* the demon does, the player doesn't know what the demon chooses. This makes backwards induction a little tricky; the demon can't assume that the player will make the best response to her choice, if the player doesn't know what the choice is.

But when we look at the numbers more closely, that worry dissipates. Whatever the demon does, the player is better off playing 2. So we should assume that the player will play 2. Given that the player will play 2, the demon is best off playing 2. And given that the demon is best off playing 2, the player at the initial node is best off playing *C*. So backwards induction leads to the same conclusion that ratificationist approaches do. And, as I said, this seems like an intuitive enough verdict, at least in this case.

In particular, a causal ratificationist might accept both of the bullet points above. It is obvious that they will accept the first. The co-operation option in Prisoners' Dilemma is both unratifiable and lower valued than the defection option. What's interesting is that, given one big extra assumption, they can accept the second as well.

The big extra assumption is that conditional on one's playing a strategy S, one should give probability 1 to the claim that the other player will do something that is in their best interests given that one is playing S. Let's apply that to Rock-Paper-Scissors. Conditional on playing Rock, one should give probability 1 to the proposition that the other player will play Paper. That is, one should give probability 0 to the proposition *If I were to play Rock, I would win*, while giving probability 1 to the proposition *If I were to play Scissors, I would win*. So conditional on playing Rock, the best thing to do, *from a causal perspective*, is to play Scissors.

So playing Rock is not ratifiable, and by similar reasoning, neither is playing Paper or Scissors. Does this mean nothing is ratifiable? Not at all; the mixed strategies might still be ratifiable. In particular, the mixed strategy where one plays each of the pure strategies with probability 1/3, is ratifiable. At least, it is ratifiable if we assume that causal ratificationism is the correct theory of rational choice. If one plays this mixed strategy, and the other player knows it, then every strategy the other player could play is equally valuable to them; they each have expected value 0. Given that each strategy is equally valuable, the other player could play any strategy that is rationally acceptable. Since we are assuming causal ratificationism, that means they could play any ratifiable strategy. But the only ratifiable strategy is the mixed strategy where one plays each of the pure strategies with probability 1/3. Conditional on the other player doing that, moving away from the mixed strategy has no advantages (though it also has no costs). So causal ratificationism, plus an assumption that the other player is an excellent mind reader, delivers the orthodox result.

There are other reasons to associate orthodoxy in game theory with causal ratificationism. Here, for instance, is Ken Binmore motivating the use of equilibrium concepts in a prominent game theory textbook (*Playing for Real*).

> Why should anyone care about Nash equilibria? There are at least two reasons. The first is that a game theory book can't authoratively point to a pair of strategies as the solution of a game unless it is a Nash equilibrium. Suppose, for example, that t weren't a best reply to s. [Player 2] would then reason that if [Player 1] follows the book's

> advice and plays s, then she would do better not to play t. But a book can't be authoritative on what is rational if rational people don't play as it predicts. (*Playing for Real*, 18-19)

It seems to me that this argument only makes sense if we assume some ratificationist theory of decision making. What Binmore is encouraging us to focus on here are strategies that are still rational strategies in situations where everyone believes that they are rational strategies. That's close, I think, to saying that we should only follow strategies that are best strategies conditional on being played.

There's a secondary argument Binmore is making here which I think is more misleading. The most the little argument he gives could show is that if a game has a unique solution, then that solution must be a Nash equilibrium. But it doesn't follow from that that there is anything special about Nash equilibrium as opposed, say, to strategies that are BRBRI. After all, if a game has a unique solution, each player will play a strategy that is BRBRI. And if each player has multiple BRBRI strategies, even if some of them are not part of any Nash equilibrium, it isn't clear why a book which said each BRBRI strategy was rational would be self-undermining. If I say that any pure or mixed strategy whatsoever could be rational in Rock-Paper-Scissors, and Player 1 believes me, and Player 2 knows this, Player 2 can't use that knowledge to undermine my advice.

But you might note that Binmore says there is a second reason. Here's what it is.

> Evolution provides a second reason why we should care about Nash equilibria. If the payoffs in a game correspond to how fit the players are, then adjustment processes that favour the more fit at the expense of the less fit will stop working when we get to a Nash equilibrium because all the survivors will then be as fit as it is possible to be in the circumstances. (*Playing for Real*, 19)

This seems to me like a very strong reason to care about Nash equilibrium in *repeated* games, like competition for food over time. The same is true in Rock-Paper-Scissors. It isn't true that Rock wins every time, and you shouldn't play Rock every time like Bart Simpson does. But that's because you'll give away your strategy. It doesn't show that a pure strategy of Rock is wrong in any given game of Rock-Paper-Scissors.

As we noted at the end of the last section, it is not clear that the standard representation of the payoffs in any given round of a repeated game are correct. Some strategies incur costs down the road that aren't properly reflected in the individual payoff matricies. But, as we also noticed, it is hard to do much about this. The best thing might be to just note that the payoffs aren't quite right, and look for equilibrium with respect to the not quite right payoffs.

So it isn't easy to come up with a straightforward decision-theoretic motivation for the use of equilibrium concepts in game theory. But maybe there are some less straightforward motivations that might be more successful.

9.8 Game Theory as Epistemology

Problems in game theory have a very different structure to problems in decision theory. In decision theory, we state what options are available to the agent, what states are epistemically possible and, and this is crucial, what the probabilities are of those states. Standard approaches to decision theory don't get off the ground until we have the last of those in place.

In game theory, we typically state things differently. Unless nature is to make a move, we simply state what options are available to the players, and what plays are available to each of the actors, and of course what will happen given each combination of moves. We are told that the players are rational, and that this is common knowledge, but we aren't given the probabilities of each move. Now it is true that you could regard each of the moves available to the other players as a possible state of the world. Indeed, I think it should be at least consistent to do that. But in general if you do that, you won't be left with a solvable decision puzzle, since you need to say something about the probabilities of those states/decisions.

So what game theory really offers is a model for simultaneously solving for the probability of different choices being made, and for the rational action given those choices. Indeed, given a game between two players, A and B, we typically have to solve for six distinct 'variables'.

1. A's probability that A will make various different choices.
2. A's probability that B will make various different choices.
3. A's choice.
4. B's probability that A will make various different choices.
5. B's probability that B will make various different choices.
6. B's choice.

The game theorists' method for solving for these six variables is typically some form of reflective equilibrium. A solution is acceptable iff it meets a number of equilibrium constraints. It isn't obvious that reflective equilibrium is the right method to use here, bu it isn't obviously wrong either.

Looked at this way, it seems that we should think of game theory really not as part of decision theory, but as much a part of epistemology. After all, what we're trying to do here is solve for what rationality requires the players credences to be, given some relatively weak looking constraints. We also try to solve for their decisions given these credences, but it turns out that is an easy part of the analysis; all the work is in the epistemology. So it isn't wrong to call this part of game theory 'interactive epistemology', as is often done.

What are the constraints on an equilibrium solution to a game? At least the following constraints seem plausible. All but the first are really equilibrium constraints; the first is somewhat of a foundational constraint. (Though note that since 'rational' here is analysed in terms of equilibria, even that constraint is something of an equilibrium constraint.)

- If there is a uniquely rational thing for one of the players to do, then both players must believe they will do it (with probability 1). More generally, if there is a unique rational credence for us to have, as theorists, about what A and B will do, the players must share those credences.
- 1 and 3, and 5 and 6, must be in equilibrium. In particular, if a player believes they will do something (with probability 1), then they will do it.
- 2 and 3, and 4 and 6, must be in equilibrium. A players decision must maximise expected utility given her credence distribution over the space of moves available to the other player.

That much seems relatively uncontroversial, assuming that we want to go along with the project of finding equilibria of the game. But those criteria alone are much too weak to get us near to game theoretic orthodoxy. After all, in Matching Pennies they are consistent with the following solution of the game.

- Each player believes, with probability 1, that they will play Heads.
- Each player's credence that the other player will play heads is 0.5.
- Each player plays Heads.

Every player maximises expected utility given the other player's expected move. Each player is correct about their own move. And each player treats the other player as being rational. So we have many aspects of an equilibrium solution. Yet we are a long way short of a Nash equilibrium of the game, since the outcome is one where one player deeply regrets their play. What could we do to strengthen the equilibrium conditions? Here are four proposals.

First, we could add a **truth** rule.

- Everything the players believe must be true. This puts constraints on 1, 2, 4 and 5.

This is a worthwhile enough constraint, albeit one considerably more externalist friendly than the constraints we usually use in decision theory. But it doesn't rule out the 'solution' I described here, since everything the players believe is true.

Second, we could add a **converse truth** rule.

- If something is true in virtue of the players' credences, then each player believes it.

This would rule out our 'solution'. After all, neither player believes the other player will play Heads, but both players will in fact play Heads. But in a slightly different case, the converse truth rule won't help.

- Each player believes, with probability 0.9, that they will play Heads.
- Each player's credence that the other player will play heads is 0.5.
- Each player plays Heads.

Now nothing is guaranteed by the players' beliefs about their own play. But we still don't have a Nash equilibrium. We might wonder if this is really consistent with converse truth. I think this depends on how we interpret the first clause. If we think that the first clause must mean that each player will use a randomising device to make their choice, one that has a 0.9 chance of coming up heads, then converse truth would say that each player should believe that they will use such a device. And then the Principal Principle would say that each player should have credence 0.9 that the other player will play Heads, so this isn't an equilibrium. But I think this is an overly *metaphysical* interpretation of the first clause. The players might just be uncertain about what they will play, not certain that they will use some particular chance device. So we need a stronger constraint.

Third, then, we could try a **symmetry** rule.

- Each player should have the same credences about what A will do, and each player should have the same credences about what B will do.

This will get us to Nash equilibrium. That is, the only solutions that are consistent with the above constraints, plus symmetry, are Nash equilibria of the original game. But what could possibly justify symmetry? Consider the following simple cooperative game.

Each player must pick either Heads or Tails. Each player gets a payoff of 1 if the picks are the same, and 0 if the picks are different.

What could justify the claim that each player should have the same credence that A will pick Heads? Surely A could have better insight into this! So symmetry seems like too strong a constraint, but without symmetry, I don't see how solving for our six 'variables' will inevitably point to a Nash equilibrium of the original game.

Perhaps we could motivate symmetry by deriving it from something even stronger. This is our fourth and final constraint, called *uniqueness*.

- There is a unique rational credence function given any evidence set.

Assume also that players aren't allowed, for whatever reason, to use knowledge not written in the game table. Assume further that there is common knowledge of rationality, as we usually assume. Now uniqueness will entail symmetry. And uniqueness, while controversial, is a well known philosophical theory. Moreover, symmetry plus the idea that we are simultaneously solving for the players' beliefs and actions gets us the result that players always believe that a Nash equilibrium is being played. And the correctness condition on player beliefs means that rational players will always play Nash equilibria.

So we have a couple of ways to justify the focus on equilibrium concepts in game theory. If we endorse a symmetry condition on credences, then we get the result that players should play their part of some Nash equilibrium or other. If we endorse a uniqueness condition on rational credence, we get a much stronger result, namely that in any given game there is a 'correct' Nash equilibrium that each player will play. I think the uniqueness condition gives implausibly strong results, and the symmetry condition is hard to justify without uniqueness, but both of those thoughts are somewhat speculative. The bigger conclusion from this section is that we can integrate game theory into the rest of philosophy more easily if we think of it as part of epistemology, as a figuring out of how rational players will react to a game, than as part of decision theory.

Chapter 10

Finding Equilibria

10.1 Finding Mixed Strategy Equilibria

Let's consider again the asymmetric version of Death in Damascus. We're writing it now not as a decision problem, but as a game played between the man and Death.

Game 23	Damascus	Aleppo
Damascus	1, -1	-1, 0.5
Aleppo	-1, 1	1, -1.5

Recall that Death is the Row player, and the Man is the Column player. The way we generated Game 23 was to take the basic structure of the game, which is Matching Pennies, and add in an 0.5 penalty for Man choosing Aleppo. It's an unpleasant journey from Damascus to Aleppo apparently, particularly if you fear Death is at the other end.

There is still no pure strategy equilibrium in this game. Whatever Death plays, Man would prefer to play the other. And whatever Man plays, Death wants to play it. So there couldn't be a set of pure choices that they would both be happy with given that they know the other's play.

But the mixed strategy equilibrium that we looked at for Matching Pennies isn't an equilibrium either. We'll write $\langle x, y \rangle$ for the mixed strategy of going to Damascus with probability x, and going to Aleppo with probability y. Clearly we should have $x + y = 1$, but it will make the representation easier to use two variables here, rather than just writing $\langle x, 1 - x \rangle$ for the mixed strategies.

Given that representation, we can ask whether the state where each player plays $\langle 1/2, 1/2 \rangle$ is a Nash equilibrium. And, as you might guess, it is not. You might have guessed this because the game is not symmetric, so it would be odd if the equilibrium solution to the game is symmetric. But let's prove that it isn't an equilibrium. Assume that Death plays $\langle 1/2, 1/2 \rangle$. Then Man's expected return from staying in Damascus is:

$$1/2 \times -1 + 1/2 \times 1 = 0$$

while his return from going to Aleppo is

$$1/2 \times 0.5 + 1/2 \times -1.5 = -0.5$$

So if Death plays $\langle 1/2, 1/2 \rangle$, Man is better off staying in Damascus than going to Aleppo. And if he's better off staying in Damascus that going to Aleppo, he's also better off staying in Damascus than playing some mixed strategy that gives some probability of going to Aleppo. In fact, the strategy $\langle x, y \rangle$ will have expected return $-y/2$, which is clearly worse than 0 when $y > 0$.

There's a general point here. The expected return of a mixed strategy is the weighted average of the returns of the pure strategies that make up the mixed strategy. In this example, for instance, if the expected value of staying in Damascus is d, and the expected value of going to Aleppo is a, the mixed strategy $\langle x, y \rangle$ will have expected value $xd + ya$. And since $x + y = 1$, the value of that will be strictly between a and d if $a \neq d$. On the other hand, if $a = d$, then $x + y = 1$ entails that $xd + ya = a = d$. So if $a = d$, then any mixed strategy will be just as good as any other, or indeed as either of the pure strategies. That implies that mixed strategies are candidates to be equilibrium points, since there is nothing to be gained by moving away from them.

This leads to an immediate, though somewhat counterintuitive, conclusion. Let's say we want to find strategies $\langle x_D, y_D \rangle$ for Death and $\langle x_M, y_M \rangle$ for Man that are in equilibrium. If the strategies are in equilibrium, then neither party can gain by moving away from them. And we just showed that that means that the expected return of Damascus must equal the expected return of Aleppo. So to find $\langle x_D, y_D \rangle$, we need to find values for x_D and y_D such that, given Man's values, staying in Damascus and leaving for Aleppo are equally valued. Note, and this is the slightly counterintuitive part, we don't need to look at *Death's* values. All that matters is that Death's strategy and Man's values together entail that the two options open to Man are equally valuable.

Given that Death is playing $\langle x_D, y_D \rangle$, we can work out the expected utility of Man's options fairly easily. (We'll occasionally appeal to the fact that $x_D + y_D =$

1.)

$$\begin{aligned}
U(\text{Damascus}) &= x_D \times -1 + y_D \times 1 \\
&= y_D - x_D \\
&= 1 - 2x_D \\
U(\text{Aleppo}) &= x_D \times 0.5 + y_D \times -1.5 \\
&= 0.5x_D - 1.5(1 - x_D) \\
&= 2x_D - 1.5
\end{aligned}$$

So there is equilibrium when $1 - 2x_D = 2x_D - 1.5$, i.e., when $x_D = 5/8$. So any mixed strategy equilibrium will have to have Death playing $\langle 5/8, 3/8 \rangle$.

Now let's do the same calculation for Man's strategy. Given that Man is playing $\langle x_D, y_D \rangle$, we can work out the expected utility of Death's options. (Again, we'll occasionally appeal to the fact that $x_M + y_M = 1$.)

$$\begin{aligned}
U(\text{Damascus}) &= x_M \times 1 + y_M \times -1 \\
&= x_M - y_M \\
&= 2x_M - 1 \\
U(\text{Aleppo}) &= x_M \times -1 + y_M \times 1 \\
&= y_M - x_M \\
&= 1 - 2x_M
\end{aligned}$$

So there is equilibrium when $2x_M - 1 = 1 - 2x_M$, i.e., when $x_M = 1/2$. So any mixed strategy equilibrium will have to have Man playing $\langle 1/2, 1/2 \rangle$. Indeed, we can work out that if Death plays $\langle 5/8, 3/8 \rangle$, and Man plays $\langle 1/2, 1/2 \rangle$, then any strategy for Death will have expected return 0, and any strategy for Man will have expected return of $-1/4$. So this pair is an equilibrium.

But note something very odd about what we just concluded. When we changed the payoffs for the two cities, we made it worse for *Man* to go to Aleppo. Intuitively, that should make Man more likely to stay in Damascus. But it turns out this isn't right, at least if the players play equilibrium strategies. The change to Man's payoffs doesn't change Man's strategy at all; he still plays $\langle 1/2, 1/2 \rangle$. What it does is change Death's strategy from $\langle 1/2, 1/2 \rangle$ to $\langle 5/8, 3/8 \rangle$.

Let's generalise this to a general recipe for finding equilibrium strategies in two player games with conflicting incentives. Assume we have the following very abstract form of a game:

Game 24	l	r
U	a_1, a_2	b_1, b_2
D	c_1, c_2	d_1, d_2

As usual, *R*ow chooses between *U*p and *D*own, while *C*olumn chooses between *l*eft and *r*ight. We will assume that R prefers the outcome to be on the northwest-southeast diagonal; that is, $a_1 > c_1$, and $d_1 > b_1$. And we'll assume that C prefers the other diagonal; that is, $c_2 > a_2$, and $b_2 > d_2$. We then have to find a pair of mixed strategies $\langle x_U, x_D \rangle$ and $\langle x_l, x_r \rangle$ that are in equilibrium. (We'll use x_A for the probability of playing A.)

What's crucial is that for each player, the expected value of each option is equal given what the other person plays. Let's compute them the expected value of playing U and D, given that C is playing $\langle x_l, x_r \rangle$.

$$U(U) = x_l a_1 + x_r b_1$$
$$U(D) = x_l c_1 + x_r d_1$$

We get equilibrium when these two values are equal, and $x_l + x_r = 1$. So we can solve for x_l the following way:

$$\begin{aligned}
& x_l a_1 + x_r b_1 = x_l c_1 + x_r d_1 \\
\Leftrightarrow\ & x_l a_1 - x_l c_1 = x_r d_1 - x_r b_1 \\
\Leftrightarrow\ & x_l(a_1 - c_1) = x_r(d_1 - b_1) \\
\Leftrightarrow\ & x_l \frac{a_1 - c_1}{d_1 - b_1} = x_r \\
\Leftrightarrow\ & x_l \frac{a_1 - c_1}{d_1 - b_1} = 1 - x_l \\
\Leftrightarrow\ & x_l \frac{a_1 - c_1}{d_1 - b_1} + x_l = 1 \\
\Leftrightarrow\ & x_l(\frac{a_1 - c_1}{d_1 - b_1} + 1) = 1 \\
\Leftrightarrow\ & x_l = \frac{1}{\frac{a_1 - c_1}{d_1 - b_1} + 1}
\end{aligned}$$

I won't go through all the same steps, but a similar argument shows that

$$x_U = \frac{1}{\frac{b_2 - a_2}{c_2 - d_2} + 1}$$

I'll leave it as an exercise to confirm these answers are correct by working out the expected return of U, D, l and r if these strategies are played.

The crucial take-away lesson from this discussion is that to find a mixed strategy equilibrium, we look at the interaction between one player's mixture and the other player's payoffs. The idea is to set the probability for each move in such a way that even if the other player knew this, they wouldn't be able to improve their position, since any move would be just as good for them as any other.

I've focussed on the case of a game where each player has just two moves. When there are more than two moves available, things are a little more complicated, but only a little. We no longer need it to be the case that one player's mixed strategy must make *every* other strategy the other player has equally valuable. It only has to make every strategy that is part of the other player's mixture equally valuable. Consider, for instance, what happens if we expand our asymmetric Death in Damascus game to give Man the option of shooting himself.

Game 25	Damascus	Aleppo	Shoot
Damascus	1, -1	-1, 0.5	-1, -2
Aleppo	-1, 1	1, -1.5	-1, -2

The shooting option is no good for anyone; Death doesn't get to meet Man, and Man doesn't get to live the extra day. So if Death plays $\langle 5/8, 3/8 \rangle$, that will make Damascus and Aleppo equally valuable to Man, but Shooting will still have an expected return of -2, rather than the expected return of $-1/4$ that Damascus and Aleppo have. But that's consistent with Death's strategy being part of an equilibrium, since Man's strategy will be to play $\langle 1/2, 1/2, 0 \rangle$. Since Man isn't playing Shoot, it doesn't matter that Shoot is less valuable to him, given Death's move, than the two pure strategies.

10.2 Coordination Games

Many games have multiple equilibria. In such cases, it is interesting to work through the means by which one equilibrium rather than another may end up being chosen. Consider, for instance, the following three games.

Game 26	*a*	*b*
A	1, 1	0, 0
B	0, 0	1, 1

Game 27	*a*	*b*
A	2, 1	0, 0
B	0, 0	1, 2

Game 28	*a*	*b*
A	5, 5	0, 4
B	4, 0	2, 2

In each case, both $\langle A, a\rangle$ and $\langle B, b\rangle$ are Nash equilibria.

Game 26 is a purely cooperative game; the players have exactly the same payoffs in every outcome. It is a model for some real-world games. The two players, R and C, have to meet up, and it doesn't matter where. They could either go to location A or B. If they go to the same location, they are happy; if not, not.

In such an abstract presentation of game, it is hard to see how we could select an equilibrium out of the two possibilities. In practice, it often turns out not to be so hard. Thomas Schelling (in *The Strategy of Conflict*) noted that a lot of real-life versions of this game have what he called 'focal points'. (These are sometimes called Schelling points now, in his honour.) A focal point is a point that stands out from the other options in some salient respect, and it can be expected that other players will notice that it stands out. Here's a nice example from a recent paper by Christopher Potts (Interpretive Economy, Schelling Points, and evolutionary stability).

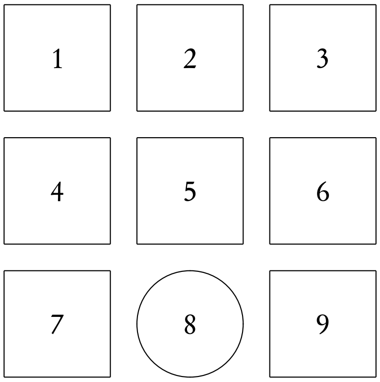
Game 29

A and *B* have to select, without communicating, one of the following nine figures. They each get a reward iff they select the same figure.

1	2	3
4	5	6
7	8	9

Game 30

A and *B* have to select, without communicating, one of the following nine figures. They each get a reward iff they select the same figure.



We could run experiments to test this, but intuitively, players will do better at Game 30 than at Game 29. That's because in Game 30, they have a focal point to select; one of the options stands out from the crowd.

Schelling tested this by asking people where they would go if they had to meet a stranger in a strange city, and had no way to communicate. The answers suggested that meetups would be much more common than you might expect. People in Paris would go to the Eiffel Tower; people in New York would go to (the information booth at) Grand Central Station, and so on. (This game may be easier if you don't actually live in the city in question. I live in New York and go to Grand Central about once a year; it would be far from the most obvious place I would select.)

Game 27 is often called 'Battle of the Sexes'. The real world example of it that is usually used is that R and C are a married couple trying to coordinate on a night out. But for whatever reason, they are at the stage where they simply have to go to one of two venues. R would prefer that they both go to A, while C would prefer that they both go to B. But the worst case scenario is that they go to different things.

Game 28 is often called 'Stag Hunt'. The idea behind the game goes back to Rousseau, although you can find similar examples in Hume. The general idea is that A is some kind of cooperative action, and B is a 'defecting' action. If players make the same selection, then they do well. If they both cooperate, they do better than if they both defect. But if they don't cooperate, it is better to defect than to cooperate.

Rousseau's example was of a group of hunters out hunting a stag, each of whom sees a hare going by. If they leave the group, they will catch the hare for sure, and have something to eat, but the group hunting the stag will be left with a much lower probability of stag catching. Hume was more explicit than Rousseau that these games come up with both two player and many player versions. Here is a nice many player version (from Ben Polak's OpenYale lectures).

> **Game 31**
>
> Everyone in the group has a choice between Investing and Not Investing. The payoff to anyone who doesn't invest is 0. If 90% or more of the group Invests, then everyone who Invests gains \$1. If less than 90% of the group Invests, then everyone who invests loses \$1.

Again, there are two Nash equilibria: everyone invests, and everyone does not invest. Hume's hypothesis was that in games like this, the cooperative move (in this case Investing) would be more likely in small groups than large groups.

10.3 Mixed Strategies in Multiple Equilibrium Games

All three of these games have mixed strategy equilibria as well as their pure strategy equilibria. For Game 26, the mixed strategy equilibrium is easy to find. Each player plays $\langle 1/2, 1/2 \rangle$. In that case the other player has an expected return of $1/2$ whatever they play, so it is an equilibrium.

Things are a little trickier in Game 27. To find the mixed strategy equilibrium there, we have to apply the lesson we learned earlier: find strategies that make the other player indifferent between their options. If R plays $\langle x, 1-x \rangle$, then C's expected return from playing a and b is:

$$\begin{aligned} U(a) &= x \times 1 + (1-x) \times 0 \\ &= x \\ U(b) &= x \times 0 + (1-x) \times 2 \\ &= 2(1-x) \end{aligned}$$

So there will be equilibrium only if $x = 2(1-x)$, i.e., only if $x = 2/3$. If that happens, C's expected return from any strategy will be $2/3$. A similar argument shows that we have equilibrium when C plays $\langle 1/3, 2/3 \rangle$, and in that case R's expected return from any strategy will be $2/3$. (Unlike in the asymmetric version of Death in Damascus, here the equilibrium strategy is to bias one's mixture towards the result that one wants.)

We apply the same methodology in Game 28. So R will play $\langle x, 1-x \rangle$, and C's expected return from the two available pure strategies is the same. Those expected returns are:

$$\begin{aligned}U(a) &= x \times 5 + (1-x) \times 0 \\ &= 5x \\ U(b) &= x \times 4 + (1-x) \times 2 \\ &= 4x + 2 - 2x \\ &= 2 + 2x\end{aligned}$$

So there will be equilibrium only if $5x = 2 + 2x$, i.e., only if $x = 2/3$. If that happens, C's expected return from any strategy will be $10/3$. Since the game is completely symmetric, a very similar argument shows that if C plays $\langle 2/3, 1/3 \rangle$, then R has the same payoff whatever she plays. So each player cooperating with probability $2/3$ is a Nash equilibrium.

In two of the cases, the mixed strategy equilibrium is worse for each player than the available pure strategy equilibria. In Game 28, the mixed equilibrium is better than the defecting equilibrium, but worse than the cooperating equilibrium. Say an equilibrium is **Pareto-preferred** to another equilibrium iff every player would prefer the first equilibrium to the second. An equilibrium is **Pareto-dominant** iff it is Pareto-preferred to all other equilibria. The cooperative equilibrium is Pareto-dominant in Game 28; neither of the other games have Pareto-dominant equiilbria.

Consider each of the above games from the perspective of a player who does not know what strategy their partner will play. Given a probability distribution over the other player's moves, we can work out which strategy has the higher expected return, or whether the various strategies have the same expected returns as each other. For any strategy S, and possible strategy S' of the other player, let f be a function that maps S into the set of probabilities x such that if the other player plays S' with probability x, S has the highest expected return. In two strategy games, we can remove the relativisation to S', since for the purposes we'll go on to, it doesn't matter which S' we use. In somewhat formal language, $f(S)$ is the **basin of attraction** for S; it is the range of probability functions that points towards S being played.

Let m be the usual (Lesbegue) measure over intervals; all that you need to know about this measure is that for any interval $[x, y]$, $m([x, y]) = y - x$; i.e., it maps a continuous interval onto its length. Say $m(f(S))$ is the risk-attractiveness of S. Intuitively, this is a measure of how big a range of probability distributions over the other person's play is compatible with S being the best thing to play.

Say that a strategy S is risk-preferred iff $m(f(S))$ is larger than $m(f(S^*))$ for any alternative strategy S^* available to that agent. Say that an equilibrium is **risk-dominant** iff it consists of risk-preferred strategies for all players.

For simple two-player, two-option games like we've been looking at, all of this can be simplified a lot. An equilibrium is risk-dominant iff each of the moves in it are moves the players would make if they assigned probability 1/2 to each of the possible moves the other player could make.

Neither Game 26 nor Game 27 have a risk-dominant equilibrium. An asymmetric version of Game 26, such as this game, does.

Game 32	*a*	*b*
A	2, 2	0, 0
B	0, 0	1, 1

In this case $\langle A, a\rangle$ is both Pareto-dominant and risk-dominant. Given that, we might expect that both players would choose A. (What if an option is Pareto-dominated, and risk-dominated, but focal? Should it be chosen then? Perhaps; the Eiffel Tower isn't easy to get to, or particularly pleasant once you're there, but seems to be chosen in the Paris version of Schelling's meetup game because of its focality.)

The real interest of risk-dominance comes from Game 28, the Stag Hunt. In that game, cooperating is Pareto-dominant, but defecting is risk-dominant. The general class of Stag Hunt games is sometimes defined as the class of two-player, two-option games with two equilibria, one of which is Pareto-dominant, and the other of which is risk-dominant. By that definition, the most extreme Stag Hunt is a game we discussed earlier in the context of deleting weakly dominated strategies.

Game 11	*l*	*r*
T	1, 1	100, 0
B	0, 100	100, 100

The $\langle B, r\rangle$ equilibrium is obviously Pareto-dominant. But the $\langle T, l\rangle$ is *extremely* risk-dominant. Any probability distribution at all over what the other player might do, except for the distribution that assigns probability 1 to B/r, makes it better to play one's part of $\langle T, l\rangle$ than one's part of $\langle B, r\rangle$.

I won't go through all the variations here, in large part because I don't understand them all, but there are a number of ways of modifying the Stag Hunt so that risk-dominant strategies are preferred. Some of these are evolutionary; given certain rules about what kinds of mutations are possible, risk-dominant strategies will invariable evolve more successfully than Pareto-dominant strategies. And some of these involve uncertainty; given some uncertainty about the payoffs available to other players, risk-dominant strategies may be uniquely rational. But going the details of these results is beyond the scope of these notes.

10.4 Value of Communication

In all the games we've discussed to date, we have assumed that the players are not able to communicate before making their choices. Or, equivalently, we've assumed that the payoff structure is what it is after communication has taken place. If we relax that assumption, we need to think a bit about the kind of speech acts that can happen in communication.

Imagine that R and C are playing a Prisoners' Dilemma. There are two importantly different types of communication that might take place before play. First, they might promise really sincerely that they won't defect. Second, they might come to some arrangement whereby the person who defects will incur some costs. This could be by signing a contract promising to pay the other in the event of defection. Or it could be by one player making a plausible threat to punish the other for defection.

The second kind of communication can change the kind of game the players are playing. The first kind does not, at least not if the players do not regard promise breaking as a bad thing. That's because the second kind of communication, but not the first, can change the payoff structure the players face. If R and C each have to pay in the event of defecting, it might be that defecting no longer dominates cooperating, so the game is not really a Prisoners' Dilemma. But if they merely say that they will cooperate, and there is no cost to breaking their word, then the game still is a Prisoners' Dilemma.

Call any communication that does not change the payoff matrix **cheap talk**. In Prisoners' Dilemma, cheap talk seems to be useless. If the players are both rational, they will still both defect.

But it isn't always the case that cheap talk is useless. In a pure coordination game, like Game 29, cheap talk can be very useful. If a player says that they will play 7, then each player can be motivated to play 7 even if they have no interest in honesty. More precisely, assume that the hearer initially thinks that the speaker is just as likely to be lying as telling the truth when she says something about what she will do. So before she thinks too hard about it, she gives credence 0.5 to the speaker actually playing 7 when she says she will do so. But if there's a 50% chance the speaker will play 7, then it seems better to play 7 than anything else. In the absence of other information, the chance that the hearer will win when playing some number other than 7 will be much less than 50%; around 6% if she has equal credence in the speaker playing each of the other options. So the hearer should play 7. But if the speaker can reason through this, then she will play 7 as well. So her statement will be self-enforcing; by making the statement she gives herself a reason to make it true, even beyond any intrinsic value she assigns to being honest.

There is one step in that reasoning that goes by particularly fast. Just because we think it is in general just as likely that a speaker is lying as telling the truth, doesn't mean that we should think those things are equally likely *on this occasion*. If the speaker has a particular incentive to lie on this occasion, the fact that they are a truth-teller half the time is irrelevant. But in Game 29, they have no such incentive. In fact, they have an incentive to tell the truth, since truth-telling is the natural way to a good outcome for them in the game.

But this consideration is a problem in Battle of the Sexes. Assume that R says, "I'm going to play A, whatever you say you'll do." If C believes this, then she has a reason to play a, and that means R has a reason to do what she says. So you might think that Game 27 is like Game 29 as a game in which cheap talk makes a difference. But in fact the reasoning that we used in Game 29 breaks down a little. Here R has an incentive to make this speech independently of what they are planning to do. Unless we think R has a general policy of truth-telling, it seems speeches like this should be discounted, since R's incentive to talk this way is independent of how the plan to play. And if R has a general policy of truth-telling, a policy they regard it as costly to break, this isn't really a case of cheap talk.

The same analysis seems to apply with even greater force in Game 28. There, R wants C to play a, whatever R is planning on playing. So she wants to give C a reason to play a. And saying that she'll play A would be, if believed, such a reason. So it seems we have a simpler explanation for why R says what she says, independent of what R plans to do. So I suspect that in both Game 27 and Game 28, this kind of cheap talk (i.e., solemn declarations of what one will play) is worthless.

But that's not all we can do when we communicate. We can also introduce new options. (Just whether this should be called genuinely cheap talk is perhaps a little dubious, but it seems to me to fall under the same general heading.) Assume that we modify Game 27 by allowing the two players to see the result of a fair coin flip. We introduce a new option for them each to play, which is to do A if the coin lands heads, and B if the coin lands tails. Call this option Z. The new game table looks like this. (Note that many of the payoffs are *expected* payoffs, not guarantees.)

Game 33	a	b	z
A	2, 1	0, 0	1, 0.5
B	0, 0	1, 2	0.5, 1
Z	1, 0.5	0.5, 1	1.5, 1.5

Note that $\langle Z, z\rangle$ is an equilibrium of the game. Indeed, it is a better equilibrium by far than the mixed strategy equilibrium that left each player with an expected return of 2/3. Not surprisingly, this kind of result is one that players with a chance to communicate often end up at.

Assume that R thinks that C will play A, B, z with probabilities $x, y, 1-x-y$. Then R's expected returns for her three strategies are:

$$\begin{aligned}
U(A) &= 2x + 0y + 1(1-x-y) \\
&= 1 + x - y \\
U(B) &= 0x + 1y + 0.5(1-x-y) \\
&= 0.5 - 0.5x + 0.5y \\
U(Z) &= 1x + 0.5y + 1.5(1-x-y) \\
&= 1.5 - 0.5x - y
\end{aligned}$$

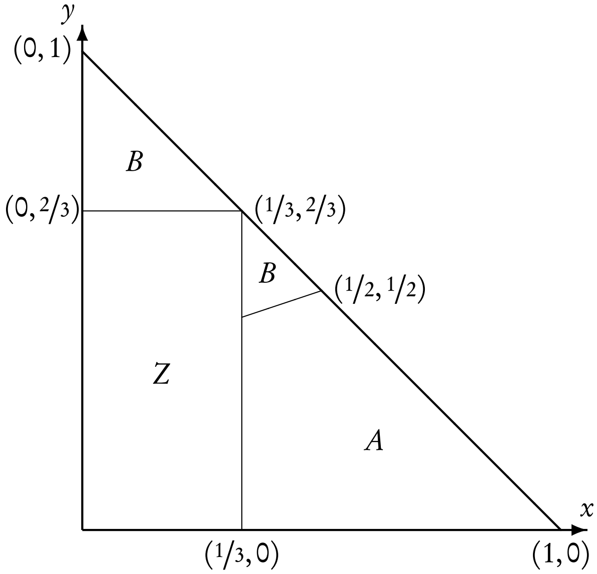
A little algebra gives us the following inequalities.

$$\begin{aligned}
U(A) > U(Z) &\Leftrightarrow 3x > 1 \\
U(A) > U(B) &\Leftrightarrow x > 3y - 1 \\
U(B) > U(Z) &\Leftrightarrow 3y > 2
\end{aligned}$$

Putting these together we get the following results:

A is best iff $x > 1/3$ and $x > 3y - 1$
B is best iff $y > 2/3$ or ($x > 1/3$ and $3y - 1 > x$)
Z is best iff $x < 1/3$ and $y < 2/3$

Here is a graph showing where each of the three options has the highest utility.



A little geometry reveals that the area of the large rectangle where Z is best is 2/9, the two triangles where B is best have area 1/18 amd 1/54 each, summing to 2/27, and the remaining area in the triangle, the odd-shaped area where A is best, is therefore 1/6. In other words, the largest region is the one where X is best. And that means, by definition, that Z is risk-preferred for R. A similar computation shows that z is risk-preferred for C. So we get the result that the newly available equilibrium is a risk-dominant equilibrium.

I'll leave the option of extending this analysis to Stag Hunt as an exercise for the interested reader.

Chapter 11

Refining Equilibria

In many games there are multiple Nash equilibria. Some of these equilibria seem highly unreasonable. For instance, the mixed equilibria in Battle of the Sexes has a lower payout to both players than either of the pure equilibria. In some coordination games, as we saw earlier, there are very poor outcomes that are still equilibria. For instance, in this game both *Aa* and *Bb* are equilibria, but we would expect players to play *Aa*, not *Bb*.

Game 34	a	b
A	1000, 1000	0, 0
B	0, 0	1, 1

Considerations such as these have pushed many game theorists to develop **refinements** of the concept of Nash equilibrium. A refinement is an equilibrium concept that is stronger than Nash equilibrium. We have already seen a couple of refinements of Nash equilibrium in the discussion of coordination games. Here are two such refinements. (The first name is common in the literature, the second I made up because there isn't a standard name.)

Pareto Nash equilibrium is a Nash equilibrium that Pareto dominates every other Nash equilibrium.

Risk Nash equilibrium is a Nash equilibrium that risk dominates every other equilibrium.

We could also consider weak versions of these concepts. A weak Pareto Nash equilibrium is a Nash equilibrium that is not Pareto-dominated by any other equilibrium, and a weak risk Nash equilibrium is a Nash equilibrium that is not risk dominated by any other Nash equilibrium. Since we've already spent a fair amount of time on these concepts when we discussed coordination games, I won't go on more about them here.

We can also generate refinements of Nash equilibrium by conjoining dominance conditions to the definition of Nash equilibrium. For instance, the following two definitions are of concepts strictly stronger than Nash equilibrium.

Nash + Weak Dominance A Nash equilibrium that is not weakly dominated by another strategy.

Nash + Iterated Weak Dominance A Nash equilibrium that is not deleted by (some process of) iterative deletion of weakly dominated strategies.

In the following game, all three of Aa, Bb and Cc are Nash equilibria, but Cc is weakly dominated, and Bb does not survive iterated deletion of weakly dominated strategies.

Game 35	a	b	c
A	3, 3	2, 0	0, 0
B	0, 2	2, 2	1, 0
C	0, 0	0, 1	1, 1

But most refinements don't come from simply conjoining an independently motivated condition onto Nash equilibrium. We'll start with the most significant refinement, subgame perfect equilibrium.

11.1 Subgame Perfect Equilibrium

Consider the following little game. The two players, call them Player I and Player II, have a choice between two options, call them Good and Bad. The game is a sequential move game; first Player I moves then Player II moves. Each player gets 1 if they choose Good and 0 if they choose Bad. We will refer to this game as **Game 36**. Here is its game tree.

A strategy for Player I in Game 36 is just a choice of one option, Good or Bad. A strategy for Player II is a little more complicated. She has to choose both what to do if Player I chooses Good, and what to do if Player II chooses Bad. We'll write her strategy as $\alpha\beta$, where α is what she does if Player I chooses Good, and β is what she does if Player II chooses Bad. (We will often use this kind of notation in what follows. Wherever it is potentially ambiguous, I'll try to explain it. But the notation is very common in works on game theory, and it is worth knowing.)

The most obvious Nash equilibrium of the game is that Player I chooses Good, and Player II chooses Good whatever Player I does. But there is another

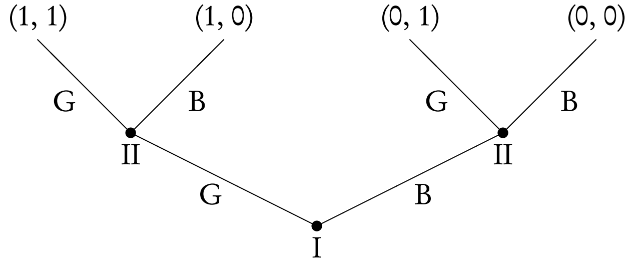


Figure 11.1: Game 36

Nash equilibrium, as looking at the strategic form of the game reveals. We'll put Player I on the row and Player II on the column. (We'll also do this from now on unless there is a reason to do otherwise.)

Game 36	gg	gb	bg	bb
G	1, 1	**1, 1**	1, 0	1, 0
B	0, 1	0, 0	0, 1	0, 0

Look at the cell that I've bolded, where Player I plays Good, and Player II plays Good, Bad. That's a Nash equilibrium. Neither player can improve their outcome, given what the other player plays. But it is a very odd Nash equilibrium. It would be very odd for Player II to play this, since it risks getting 0 when they can guarantee getting 1.

It's true that Good, Bad is weakly dominated by Good, Good. But as we've already seen, and as we'll see in very similar examples to this soon, there are dangers in throwing out *all* weakly dominated strategies. Many people think that there is something else wrong with what Player II does here.

Consider the sub-game that starts with the right-hand decision node for Player II. That isn't a very interesting game; Player I has no choices, and Player II simply has a choice between 1 and 0. But it is a game. And note that Good, Bad is not a Nash equilibrium of that game. Indeed, it is a *strictly* dominated strategy in that game, since it involves taking 0 when 1 is freely available.

Say that a strategy pair is a **subgame perfect equilibrium** when it is a Nash equilibrium, and it is a Nash equilibrium of every sub-game of the game. The pair Good and Good, Bad is not subgame perfect, since it is not a Nash equilibrium of the right-hand subgame.

When we solve extensive form games by backwards induction, we not only find Nash equilibria, but subgame perfect equilibria. Solving this game by backwards induction would reveal that Player II would choose Good wherever she ends up, and then Player I will play Good at the first move. And the only subgame perfect equilibrium of the game is that Player I plays Good, and Player II plays Good, Good.

11.2 Forward Induction

In motivating subgame perfect equilibrium, we use the idea that players will suppose that future moves will be rational. We can also develop constraints based around the idea that past moves were rational. This kind of reasoning is called **forward induction**. A clear, and relatively uncontroversial, use of it is in this game by Cho and Kreps's paper "Signalling Games and Stable Equilibria" (QJE, 1987). We'll return to that paper in more detail below, but first we'll discuss **Game 37**.

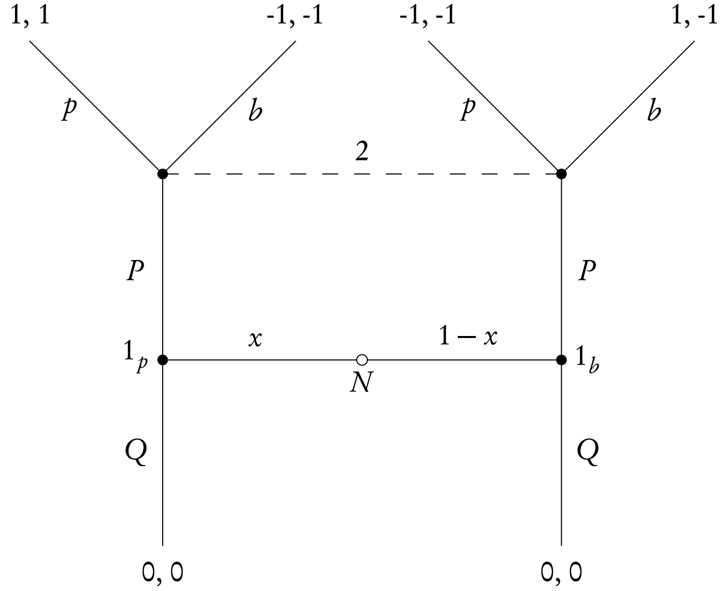


Figure 11.2: Game 37

This game takes a little more interpreting than some we've looked at previously. We use the convention that an empty disc indicates the initial node of the game. In this case, as in a few games we'll look at below, it is in the middle of the tree. The first move is made by *N*ature, denoted by N. Nature makes Player 1 playful with probability x, or bashful, with probability $1-x$. Player 1's personality type is revealed to Player 1, but not to Player 2. If she's playful, she moves to the node marked 1_p, if bashful to the node marked 1_b. Either way she has a choice about whether to play a guessing game with Player 2, where Player 2 has to guess her personality type. Player 1's payouts are as follows:

- If she doesn't play the guessing game, she gets 0.
- If she's playful, she gets 1 if Player 2 guesses correctly, and -1 if Player 2 guesses incorrectly.
- If she's bashful, she gets -1 either way.

Player 2's payouts are a little simpler.

- If the guessing game isn't played, she gets 0.
- If it is played and she guesses correctly, she gets 1.
- If it is played and she guesses wrongly, she gets -1.

The horizontal dashed line at the top indicates that if one of the upper nodes is reached, Player 2 doesn't know which node she is at. So we can't simply apply backward induction. Indeed, there aren't any subgames of this game, since there are no nodes that are neither initial nor terminal such that when they are reached, both players know they are there.

Player 1 has four possible strategies. She has to decide whether to *P*lay or *Q*uit both for when she is playful and when she is bashful. We'll write a strategy $\alpha\beta$ as the strategy of playing α if playful, and β if bashful. (We're going to repeat a lot of this notation when we get to signalling games, but it is worthwhile going over it a few times to be maximally clear.) Player 2 only has one choice and two possible strategies: if she gets to guess, she can guess *p*layful or *b*ashful. If we set out the strategic form of the game, we get the following expected payouts. (It's worth checking that you understand why these are the expected payouts for each strategy.)

Game 37	p	b
PP	$2x-1, 2x-1$	$-1, 1-2x$
PQ	x, x	$-x, -x$
QP	$x-1, x-1$	$x-1, 1-x$
QQ	$0, 0$	$0, 0$

Assuming $0 < x < 1$, it is easy to see that there are two pure Nash equilibria here: $\langle PQ, p\rangle$ and $\langle QQ, r\rangle$. But there is something very odd about the second equilibrium. Assume that both players are rational, and Player 2 actually gets to play. If Player 1 is bashful, then Quitting dominates *P*laying. So a rational bashful Player 1 wouldn't give Player 2 a chance to move. So if Player 2 gets a chance to move, Player 1 must be playful. And if Player 1 is playful, the best move for Player 2 is p. So by forward induction reasoning, Player 2 should play p. Moreover, Player 1 can figure all this out, so by backward induction reasoning she should play her best response to p, namely PQ.

We'll look at reasons for being sceptical of forward induction reasoning, or at least of some notable applications of it, next time. But at least in this case, it seems to deliver the right verdict: the players should get to the $\langle PQ, p\rangle$ equilibrium, not the $\langle QQ, r\rangle$ equilibrium.

11.3 Perfect Bayesian Equilibrium

The core idea behind subgame perfect equilibrium was that we wanted to eliminate equilibria that relied on 'incredible threats'. That is, there are some equilibria such that the first player makes a certain move only because if they make a different move, the second player to move would, on their current strategy, do something that makes things worse for both players. But there's no reason to think that the second player would actually do that.

The same kind of consideration can arise in games where there aren't any subgames. For instance, consider the following game, which we'll call **Game 38**.

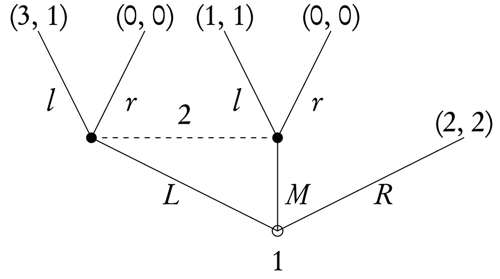


Figure 11.3: Game 38

The game here is a little different to one we've seen so far. First Player 1 makes a move, either L, M or R. If her move is R, the game ends, with the (2, 2) payout. If she moves L or M, the game continues with a move by Player 2. But crucially, Player 2 does not know what move Player 1 made, so she does not know which node she is at. So there isn't a subgame here to start. Player 2 chooses l or r, and then the game ends.

There are two Nash equilibria to Game 38. The obvious equilibrium is Ll. In that equilibrium Player 1 gets her maximal payout, and Player 2 gets as much as she can get given that the rightmost node is unobtainable. But there is another Nash equilibrium available, namely Rr. In that equilibrium, Player 2 gets her maximal payout, and Player 1 gets the most she can get given that Player 2 is playing r. But note what a strange equilibrium it is. It relies on the idea that Player 2 would play r were she to be given a move. But that is absurd. Once she gets a move, she has a straight choice between 1, if she plays l, and 0, if she plays r. So obviously she'll play l.

This is just like the examples that we used to motivate subgame perfect equilibrium, but that doesn't help us here. So we need a new concept. The core idea is that each player should be modelled as a Bayesian expected utility maximiser. More formally, the following constraints are put on players.

1. At each point in the game, each player has a probability distribution over where they are in the game. These probability distributions are correct about the other players' actions. That is, if a player is playing a strategy S, everyone has probability 1 that they are playing S. If S is a mixed strategy, this might involve having probabilities between 0 and 1 in propositions about which move the other player will make, but players have correct beliefs about other players' strategies.
2. No matter which node is reach, each player is disposed to maximise expected utility on arrival at that node.
3. When a player had a positive probability of arriving at a node, on reaching that node they update by conditionalisation.
4. When a player gave 0 probability to reaching a node (e.g., because the equilibrium being played did not include that node), they have some disposition or other to form a set of consistent beliefs at that node.

The last constraint is very weak, but it does enough to eliminate the equilibrium Rr. The constraint implies that when Player 2 moves, she must have some probability distribution Pr such that there's an x such that $\Pr(L) = x$ and $\Pr(M) = 1 - x$. Whatever value x takes, the expected utility of l given Pr is 1, and the expected utility of r is 0. So being disposed to play r violates the second condition. So Rr is not a Perfect Bayesian equilibrium.

It's true, but I'm not going to prove it, that all Perfect Bayesian equilibria are Nash equilibria. It's also true that the converse does not hold, and this we have proved; Game 38 is an example.

11.4 Signalling Games

Concepts like Perfect Bayesian equilibrium are useful for the broad class of games known as signalling games. In a signalling game, Player 1 gets some information that is hidden from player 2. Many applications of these games involve the information being information about Player 1's nature, so the information that Player 1 gets is referred to as her *type*. But that's inessential; what is essential is that only one player gets this information. Player 1 then has a choice of move to make. There is a small loss of generality here, but we'll restrict our attention to games where Player 1's choices are independent of her type, i.e., of the information she receives. Player 2 sees Player 1's move (but not, remember, her type) and then has a choice of her own to make. Again with a small loss of generality, we'll restrict attention to cases where Player 2's available moves are independent of what Player 1 does. We'll start, in game **Game 39** with a signalling game where the parties' interests are perfectly aligned.

As above, we use an empty disc to signal the initial node of the game tree. In this case, it is the node in the centre of the tree. The first move is made by Nature, again denoted as N. Nature assigns a type to Player 1; that is, she makes some proposition true, and reveals it to Player 1. Call that proposition q. We'll say that Nature moves left if q is true, and right otherwise. We assume the probability (in some good sense of probability) of q is p, and this is known before the start of the game. After Nature moves, Player 1 has to choose Up or Down. Player 2 is shown Player 1's choice, but not Nature's move. That's the effect of the horizontal dashed lines. If the game reaches one of the upper nodes, Player 2 doesn't know which one it is, and if it reaches one of the lower nodes, again Player 2 doesn't know which it is. Then Player 2 has a make a choice, here simply denoted as $l$eft or $r$ight.

In any game of this form, each player has a choice of four strategies. Player 1 has to choose what to do if q is true, and what to do if q is false. We'll write $\alpha\beta$ for the strategy of doing α if q, and β if $\neg q$. Since α could be U or D, and β could be U or D, there are four possible strategies. Player 2 has to choose what to do if U is played, and what to do if D is played. We'll write $\gamma\delta$ for the strategy of doing γ if U is played, and δ if D is played. Again, there are four possible choices. (If Player 2 knew what move Nature had made, there would be four degrees of freedom in her strategy choice, so she'd have 16 possible strategies. Fortunately, she doesn't have that much flexibility!)

Any game of this broad form is a signalling game. Signalling games differ in (a) the interpretation of the moves, and (b) the payoffs. Game 39 has a very simple payoff structure. Both players get 1 if Player 2 moves l iff q, and 0 otherwise. If we think of l as the formation of a belief that q, and r as the formation

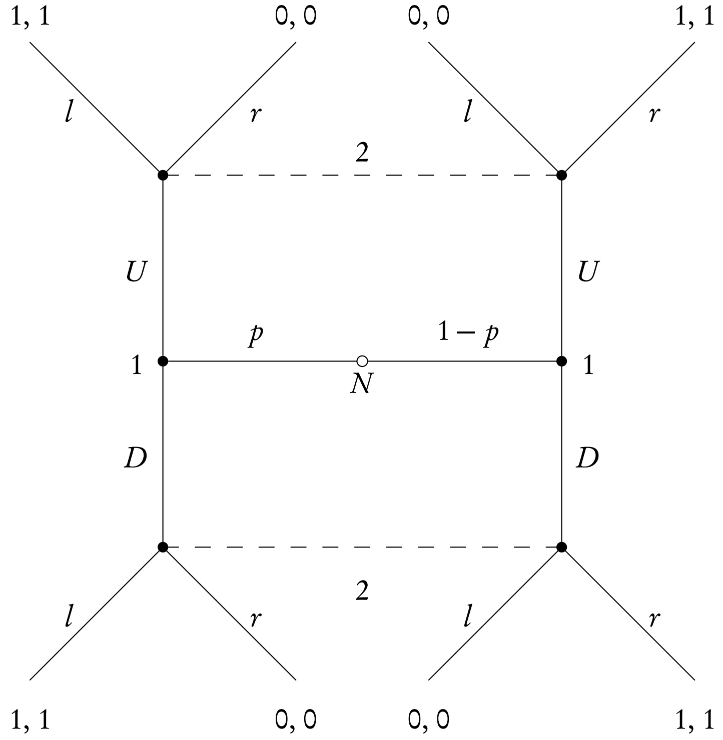


Figure 11.4: Game 39

of the opposite belief, this becomes a simple communication game. The players get a payoff iff Player 2 ends up with a true belief about q, so Player 1 is trying to communicate to Player 2 whether q is true. This kind of simple communication game was Lewis used in *Convention* to show that game theoretic approaches could be fruitful in the study of meaning. The game is perfectly symmetric if $p = 1/2$; so as to introduce some asymmetries, I'll work through the case where $p = 3/5$.

Game 39 has a dizzying array of Nash equilibria, even given this asymmetry introducing assumption. They include the following:

- There are two **separating** equilibria, where what Player 1 does depends on what Nature does. These are $\langle UD, lr\rangle$ and $\langle DU, rl\rangle$. These are rather nice equilibria; both players are guaranteed to get their maximal payout.
- There are two **pooling** equilibria, where what Player 1 does is independent

of what Nature does. These are $\langle UU, ll\rangle$ and $\langle DD, ll\rangle$. Given that she gets no information from Player 1, Player 2 may as well guess. And since $\Pr(q) > 1/2$, she should guess that q is true; i.e., she should play l. And given that Player 2 is going to guess, Player 1 may as well play anything. So these are also equilibria.

- And there are some **babbling** equilibria. For instance, there is the equilibrium where Player 1 plays either UU or DD with some probability r, and Player 2 plays ll.

Unfortunately, these are all Perfect Bayesian equilibria too. For the separating and babbling equilibria, it is easy to see that conditionalising on what Player 1 plays leads to Player 2 maximising expected utility by playing her part of the equilibrium. And for the pooling equilibria, as long as the probability of q stays above $1/2$ in any 'off-the-equilibrium-path' play (e.g., conditional on D in the $\langle UU, ll\rangle$ equilibrium), Player 2 maximises expected utility at every node.

That doesn't mean we have nothing to say. For one thing, the separating equilibria are Pareto-dominant in the sense that both players do better on those equilibria than they do on any other. So that's a non-coincidental reason to think that they will be the equilibria that arise. There are other refinements on Perfect Bayesian equilibria that are more narrowly tailored to signalling games. We'll introduce them by looking at a couple of famous signalling games.

Economists have been interested for several decades in games that give college a signalling function. So consider the following variant of the signalling game, **Game 40**. It has the following intended interpretation:

- Player 1 is a student and potential worker, Player 2 is an employer.
- The student is either bright or dull with probability p of being bright. Nature reveals the type to the student, but only the probability to the employer, so q is that the student is bright.
- The student has the choice of going to college (U) or the beach (D).
- The employer has the choice of hiring the student (l) or rejecting them (r).

In Game 40 we make the following extra assumptions about the payoffs.

- Being hired is worth 4 to the student.
- Going to college rather than the beach costs the bright student 1, and the dull student 5, since college is much harder work for dullards.
- The employer gets no benefit from hiring college students as such.
- Hiring a bright student pays the employer 1, hiring a dull student costs the employer 1, and rejections have no cost or benefit.

The resulting game tree is shown in 11.5. In this game, dull students never prefer to go to college, since even the lure of a job doesn't make up for the pain of actually having to study. So a rational strategy for Player 1 will never be of the form αU, since for dull students, college is a dominated option, being dominated by αD. But whether bright students should go to college is a trickier question. That is, it is trickier to say whether the right strategy for Player 1 is UD or DD, which are the two strategies consistent with eliminating the strictly dominated strategies. (Remember that strictly dominated strategies cannot be part of a Nash equilibrium.)

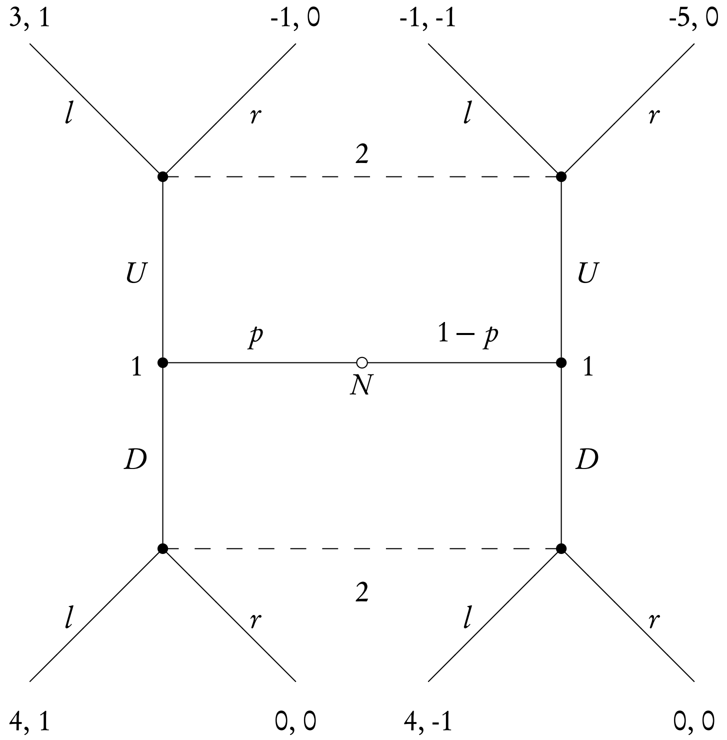


Figure 11.5: Game 40

First, consider the case where $p < 1/2$. In that case, if Player 1 plays DD, then Player 2 gets $2p - 1$ from playing either ll or rl, and 0 from playing lr or rr. So she should play one of the latter two options. If she plays rr, then DD is a best response, since when employers aren't going to hire, students prefer to go to the beach. So there is one **pooling** equilibrium, namely $\langle DD, rr \rangle$. But what if Player 2 plays lr. Then Player 1's best response is UD, since bright students prefer college conditional on it leading to a job. So there is also a **separating** equilibrium, namely $\langle UD, lr \rangle$. The employer prefers that equilibrium, since her payoff is now p rather than 0. And students prefer it, since their payoff is $3p$ rather than 0. So if we assume that people will end up at Pareto-dominant outcomes, we have reason to think that bright students, and only bright students, will go to college, and employers will hire college students, and only college students. And all this is true despite there being no advantage whatsoever to going to college in terms of how good an employee one will be.

Especially in popular treatments of the case, the existence of this kind of model can be used to motivate the idea that college *only* plays a signalling role. That is, some people argue that in the real world college does not make students more economically valuable, and the fact that college graduates have better employment rates, and better pay, can be explained by the signalling function of college. For what it's worth, I highly doubt that is the case. The wage premium one gets for going to college tends to *increase* as one gets further removed from college, although the further out from college you get, the less important a signal college participation is. One can try to explain this fact too on a pure signalling model of college's value, but frankly I think the assumptions needed to do so are heroic. The model is cute, but not really a picture of how actual college works.

So far we assumed that $p < 1/2$. If we drop that assumption, and assume instead that $p > 1/2$, the case becomes more complicated. Now if Player 1 plays DD, i.e., goes to the beach no matter what, Player 2's best response is still to hire them. But note that now a very odd equilibrium becomes available. The pair $\langle DD, rl \rangle$ is a Nash equilibrium, and, with the right assumptions, a Perfect Bayesian equilibrium. This pair says that Player 1 goes to the beach whatever her type, and Player 2 hires only beach goers.

This is a very odd strategy for Player 2 to adopt, but it is a little hard to say just why it is odd. It is clearly a Nash equilibrium. Given that Player 2 is playing rl, then clearly beach-going dominates college-going. And given that Player 1 is playing DD, playing rl gets as good a return as is available to Player 2, i.e., $2p-1$. Could it also be a Perfect Bayesian equilibrium? It could, provided Player 2 has a rather odd update policy. Say that Player 2 thinks that if someone goes to college, they are a dullard with probability greater than $1/2$. That's consistent with what we've said; given that Player 1 is playing DD, the probability that a student goes to college is 0. So the conditional probability that a college-goer is bright is left open, and can be anything one likes in Perfect Bayesian equilibrium. So if Player 2 sets it to be, say, 0, then the rational reaction is to play rl.

But now note what an odd update strategy this is for Player 2. She has to assume that if someone deviates from the *DD* strategy, it is someone for whom the deviation is strictly dominated. Well, perhaps it isn't crazy to assume that someone who would deviate from an equilibrium isn't very bright, so maybe this isn't the oddest assumption in this particular context. But a few economists and game theorists have thought that we can put more substantive constraints on probabilities conditional on 'off-the-equilibrium-path' behaviour. One such constraint, is, roughly, that deviation shouldn't lead to playing a dominated strategy. This is the "**intuitive criterion**" of Cho and Kreps. In this game, all the criterion rules out is the odd pair $\langle DD, rl \rangle$. It doesn't rule out the very similiar pair $\langle DD, ll \rangle$. But the intuitive criterion makes more substantial constraints in other games. We'll close the discussion of signalling games with such an example, and a more careful statement of the criterion.

The tree in 11.6 represents **Game 41**, which is also a guessing game. The usual statement of it involves all sorts of unfortunate stereotypes about the type of men who have quiche and/or beer for breakfast, and I'm underwhelmed by it. So I'll run with an example that relies on different stereotypes.

A North American tourist is in a bar. 60% of the North Americans who pass through that bar are from Canada, the other 40% are from the US. (This is a little implausible for most parts of the world, but maybe it is very cold climate bar. In Cho and Kreps' version the split is 90/10 not 60/40, but all that matters is which side of 50/50 it is.) The tourist, call her Player 1, knows her nationality, although the barman doesn't. The tourist can ask for the bar TV to be turned to hockey or to baseball, and she knows once she does that the barman will guess at her nationality. (The barman might also try to rely on her accent, but she has a fairly neutral upper-Midwest/central Canada accent.) Here are the tourist's preferences.

- If she is American, she has a (weak) preference for watching baseball rather than hockey.
- If she is Canadian, she has a (weak) preference for watching hockey rather than baseball.
- Either way, she has a strong preference for being thought of as Canadian rather than American. This preference is considerably stronger than her preference over which sport to watch.

The barman's preferences are simpler; he prefers to make true guesses to false guesses about the tourist's nationality. All of this is common knowledge. So the decision tree is İn this tree, *B* means asking for baseball, *H* means asking for hockey, *a* means guessing the tourist is from the USA, *c* means guessing she is from Canada.

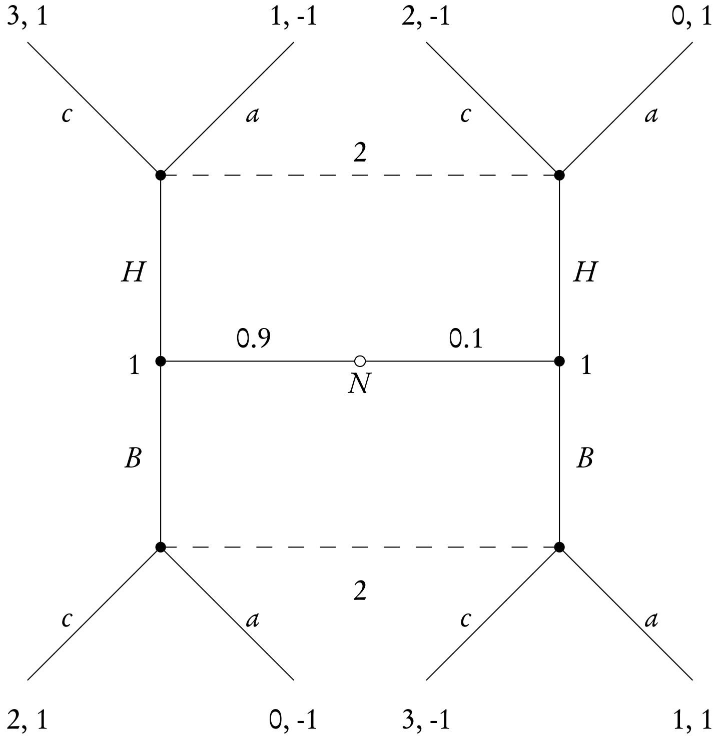


Figure 11.6: Game 41

There are two Nash equilibria for this game. One is $\langle HH, ca\rangle$; everyone asks for hockey, and the barman guesses Canadian if hockey, American if baseball. It isn't too hard to check this is an equilibrium. The Canadian gets her best outcome, so it is clearly an equilibrium for her. The American gets the second-best outcome, but asking for baseball would lead to a worse outcome. And given that everyone asks for hockey, the best the barman can do is go with the prior probabilities, and that means guessing Canadian. It is easy to see how to extend this to a perfect Bayesian equilibrium; simply posit that conditional on baseball being asked for, the probability that the tourist is American is greater than 1/2.

The other equilibrium is rather odder. It is $\langle BB, ac\rangle$; everyone asks for baseball, and the barman guesses American if hockey, Canadian if baseball. Again, it isn't too hard to see how it is Nash equilibrium. The Canadian would rather have hockey, but not at the cost of being thought American, so she has no incentive to defect. The American gets her best outcome, so she has no incentive to defect. And the barman does as well as he can given that he gets no information out of the request, since everyone requests baseball.

Surprisingly, this could also be a perfect Bayesian equilibrium. This requires that the barman maximise utility at every node. The tricky thing is to ensure he maxmises utility at the node where hockey is chosen. This can be done provided that conditional on hockey being chosen, the probability of American rises to above 1/2. Well, nothing we have said rules that out, so there *exists* a perfect Bayesian equilibrium. But it doesn't seem very plausible. Why would the barman adopt just *this* updating disposition?

One of the active areas of research on signalling games is the development of formal rules to rule out intuitively crazy updating dispositions like this. I'm just going to note one of the very early attempts in this area, Cho and Kreps' *Intuitive Criterion*.

Start with some folksy terminology that is not standard in the literature, but I think helpful at understanding what's going on. Let an outcome o of the game be any combination of moves by nature and the players. Call the players' moves collectively P and nature's move N. So an outcome is a pair $\langle P, N \rangle$. Divide the players into three types.

- The **happy** players in o are those such that given N, o maximises their possible returns. That is, for all possible P', their payoff in $\langle P, N \rangle$ is at least as large as their payoff in $\langle P', N \rangle$.
- The **content** players are those who are not happy, but who are such that for no strategy s' which is an alternative to their current strategy s, would they do better given what nature and the other players would do if they played s'.
- The **unhappy** players are those who are neither happy nor content.

The standard Nash equilibrium condition is that no player is unhappy. So assume we have a Nash equilibrium with only happy and content players. And assume it is also a Perfect Bayesian equilibrium. That is, for any 'off-equilibrium' outcome, each player would have some credence function such that their play maximises utility given that function.

Now add one constraint to those credence functions:

Intuitive Criterion - Weak Version

> Assume a player has probability 1 that a strategy combination P will be played. Consider their credence function conditional on another player playing s', which is different to the strategy s they play in P.

> If it is consistent with everything else the player believes that s' could be played by a player who is content with the actual outcome $\langle P, N \rangle$, then give probability 1 to s' being played by a content player.

That is, if some deviation from equilibrium happens, give probability 1 to it being one of the content players, not one of the happy players, who makes the deviation.

That's enough to rule out the odd equilibrium for the baseball-hockey game. The only way that is a Perfect Bayesian equilibrium is if the barman responds to a hockey request by *increasing* the probability that the tourist is American. But in the odd equilibrium, the American is happy, and the Canadian is merely content. So the Intuitive Criterion says that the barman should give credence 1 to the hockey requester, i.e., the deviator from equilibrium, being Canadian. And if so, the barman won't respond to a hockey request by guessing the requestor is American, so the Canadian would prefer to request hockey. And that means the outcome is no longer an equilibrium, since the Canadian is no longer even content with requesting baseball.

In games where everyone has only two options, the weak version I've given is equivalent to what Cho and Kreps offers. The official version is a little more complicated. First some terminology of mine, then their terminology next.

- A player is **happy to have played** s rather than s' if the payoff for s in o is greater than any possible outcome for s' given N.
- A player is **content to have played** s rather than s' if they are not happy to have played s rather than s', but the payoff for s in o is as great as the payoff for s' given N and the other players' strategies.
- A player is **unhappy to have played** s rather than s' if they are neither happy nor content to have played s rather than s'.

If a player is happy to have played s rather than s', and s is part of some equilibrium outcome o, then Cho and Kreps say that s' is an **equilibrium-dominated** strategy. The full version of the Intuitive Criterion is then:

> **Intuitive Criterion - Full Version**
>
> Assume a player has probability 1 that a strategy combination P will be played. Consider their credence function conditional on another player playing s', which is different to the strategy s they play in P. If it is consistent with everything else the player believes that s' could be played by a player who is merely content to have played s rather than s', then give probability 1 to s' being played by someone who is

> content to have played s, rather than someone who is happy to have played s.

This refinement matters in games where players have three or more options. It might be that a player's options are s_1, s_2 and s_3, and their type is t_1 or t_2. In the Nash equilibrium in question, they play s_1. If they are type t_1, they are happy to have played s_1 rather than s_2, but merely content to have played s_1 rather than s_3. If they are type t_2, they are happy to have played s_1 rather than s_3, but merely content to have played s_1 rather than s_2. In my terminology above, both players are content rather than happy with the outcome. So the weak version of the Intuitive Criterion wouldn't put any restrictions on what we can say about them conditional on them playing, say s_2. But the full version does say something; it says other players should assign probability 1 to the player being type t_2 conditional on them playing s_2, since the alternative is that s_2 is played by a player who is happy they played s_1 rather than s_2. Similarly, it says that conditional on the player playing s_3, other players should assign probability 1 to their being of type t_1, since the alternative is to assign positive probability to a player deviating to a strategy they are happy not to play.

The Intuitive Criterion has been the subject of an enormous literature. Google Scholar lists nearly 2000 citations for the Cho and Kreps paper alone, and similar rules are discussed in other papers. So there are arguments that it is too weak and arguments too strong, and refinements in all sorts of directions. But in the interests of getting back to more traditionally philosophical discussions, we'll leave those. What I most wanted to stress was the *form* a refinement of Perfect Bayesian equilibrium would take. It is a constraint on the conditional credences of agents conditional on some probability 0 event happening. It's interesting that there are some plausible, even intuitive constraints; sometimes it seems the economic literature has investigated rationality under 0 probability evidence more than philosophers have!

11.5 Other Types of Constraint

We don't have the time to go into other kinds of constraints in as much detail, but I wanted to quickly mention two other ways in which game theorists have attempted to restrict Nash equilibrium.

The first is sometimes called **trembling-hand equilibrium**. The idea is that we should restrict our attention to those strategies that are utility maximising given a very very high credence that the other players will play the strategies they actually play, and some positive (but low) credence that they will play each other strategy. This is, I think very important to real-world applications, since it is very implausible that we should assign probability 1 to any claim about the other players, particularly claims of the form that they will play some equilibrium rather than another. (There is a connection here to the philosophical claim that rational credences are *regular*; that is, that they assign positive probability to anything possible.)

In normal form games, the main effect of this is to rule out strategies that are weakly dominated. Remember that there are many strategies that are equilibria that are weakly dominated, since equilibrium concepts typically only require that player can't do better given some other constraint. But if we have to assign positive probability to any alternative, then the weakly dominating strategy will get a utility boost from the alternative under which it is preferable.

Things get more complicated when we put a 'trembling-hand' constraint on solutions to extensive form games. The usual idea is that players should, at each node, assign a positive probability to each deviation from equilibrium play. This can end up being a rather tight constraint, especially when combined with such ideas as subgame-perfection.

The other kind of refinement I'll briefly discuss in **evolutionary stability**. This is, as the name suggests, a concept that arose out of game-theoretic models of evolution. As such, it only really applies to symmetric games. In such a game, we can write things like $U(s,s')$, meaning the payoff that one gets for playing s when the other player is playing s'. In an asymmetric game, we'd have to also specify which 'side' one was when playing s, but there's no need for that in a symmetric game.

We then say that a strategy is evolutionarily stable iff these two conditions hold.

$$\begin{gathered}\forall t : (U(s,s) \geq U(t,s)) \\ \forall t \neq s : (U(s,s) > U(t,s) \vee U(s,s) > U(t,t))\end{gathered}$$

The idea is that a species playing s is immune to invasion if it satisfies these conditions. Any invader will play some alternative strategy t. Think then of a species playing t whose members have to play either against the existing strategy s with high probability, or against other members of their own species with low probability. The first clause says that the invader can't do better in the normal case, where they play with a dominant strategy. The second clause says that they do worse in one of the two cases that come up during the invasion, either playing the dominant strategy or playing their own kind. The effect is that the dominant strategy will do better, and the invader will die out.

For real-world applications we might often want to restrict the quantifier to biologically plausible strategies. And, of course, we'll want to be very careful with the specification of the game being played. But there are some nice applications of this concept in explanations of certain equilibrium outcomes. That's a topic for biology class though - not decision theory!

Chapter 12

Backward Induction and its Discontents

12.1 Problems with Backwards Induction

When each player only moves once, it is very plausible that each player should play their part of a subgame perfect equilibrium solution to the game. But this is less compelling when players have multiple moves. We can start to see why by looking at a game that Robert Stalnaker has used in a few places. Again, it is an extensive form game. We will call it **Game 42**.

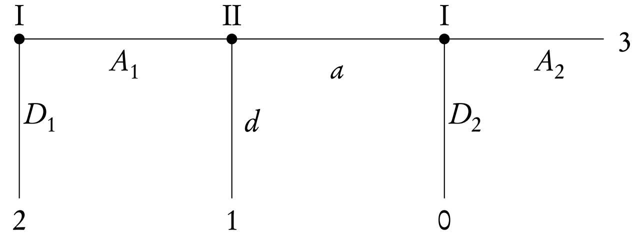


Figure 12.1: Game 42

The game starts in the upper-left corner. There are up to three moves, each of them Across or Down. As soon as one player moves Down, the game ends. It is a common interest game; each player gets the payout at the terminal node.

Since there is only one path to each decision node, a strategy merely has to consist of a set of plans for what to play if that node is reached. Alice's strategy will consist of two capitalised moves, and Bob's strategy will consist of one lowercase move.

If we apply backwards induction, we get that the unique solution of the game is $\langle A_1A_2, a\rangle$. Alice would play A_2 if it gets that far. Given that, Bob would be better off playing a than d if he gets to play. And given that, Alice is better off playing A_1 than D_1.

But there are many other Nash equilibria of the game. One of them is $\langle D_1A_2, d\rangle$. Given that Player I is playing D_1, Player II can play anything without changing her payout; it will be 2 come what may. Given that Player II is playing d, Player I is best off playing D_1 and taking 2, rather than just getting the 1 that comes from leaving Player II to make a play.

Could this equilibrium be one that rational players, who know each other to be rational, reach? Stalnaker argues that it could. Assume each player knows that the other player is rational, and is playing that strategy. Given what Player I knows, her strategy choice is clearly rational. She takes 2 rather than the 1 that she would get by playing A_1, and she is disposed to take 3 rather than 0 if she gets to the final decision node. So her actual choice is rational, and her disposition is to make another rational choice if we reach the end of the game.

Things are a little trickier for Player II. You might think it is impossible for rational Player II, who knows Player I to be rational, to move d. After all, if Player I is rational, then she'll play A_2, not D_2. And if she plays A_2, it is better to play a than d. So it looks hard to justify Player II's move. But looks can be deceiving. In fact there isn't anything wrong with Player II's move, as long as he has the right beliefs to justify it. It's very important to distinguish the following two conditionals.

- If Player I has the choice, she chooses A_2 over D_2.
- If Player I were to have the choice, she would choose A_2 over D_2.

Player II knows that the first, indicative, conditional is true. And indeed it is true. But he doesn't know that the second, subjunctive, conditional is true. After all, if Player I were to have the choice between A_2 and D_2, she would have, *irrationally*, chosen A_1 over D_1. And if she had chosen irrationally once, it's possible that she would choose irrationally again.

Here's an analogy that may help explain what's going on. The following set of beliefs is consistent.

- Any perfectly rational being gets all of the questions on their algebra exam right.
- Alice is perfectly rational.

- If Alice had got the second-hardest question on the algebra exam wrong, she would have got the hardest question on the algebra exam wrong as well.

Player II's beliefs are like that. He believes Player I is perfectly rational. He also believes that if Player I were to make an irrational move, she would continue to make irrational moves. That's consistent with belief in perfect rationality, and nothing about the game setup rules out such a belief. He also believes that playing A_1 would be irrational. That's correct, given what Player I knows about Player II. Given all those beliefs, playing d, if he had the chance, would be rational.

Stalnaker argues that many game theorists have tacitly confused indicative and subjunctive conditionals in reasoning about games like Game 42. Let's look at some other games where similar reasoning takes place.

12.2 Money Burning Game

Elchanen Ben-Porath and Eddie Dekel suggested this variant to the Battle of the Sexes game. The variation is in two steps. First, we change the payouts for the basic game to the following. (Note that I'm using a lowercase letter for one of R's options here; this is to distinguish the d of down from the D of Don't burn.)

	l	r
u	4, 1	0, 0
d	0, 0	1, 4

Then we give Player I the option of publicly burning 2 utils before playing this game. We will use D for Don't burn, and B for burn. So actually each player has four choices. Player I has to choose both D or B, then u or d. Player 2 has to choose whether to play l or r in each of the two possibilities: first, when D is played, second, when B is played. We'll write lr for the strategy of playing l if D, and r if B, and so on for the other strategies.

Game 43	ll	lr	rl	rr
Du	4, 1	4, 1	0, 0	0, 0
Dd	0, 0	0, 0	1, 4	1, 4
Bu	2, 1	-2, 0	2, 1	-2, 0
Bd	-2, 0	-1, 4	-2, 0	-1, 4

Now we can analyse this game a couple of ways. First, we can go through eliminating weakly dominated strategies. Note that Du strictly dominated Bd, so we can eliminate it. If Bd is out, then ll weakly dominates lr, and rl weakly dominates rr. So we can eliminate lr and rr. Now Bu weakly dominates Dd, relative to the remaining options, so we can eliminate it. Given that just Du and Bu remain, ll weakly dominates all options for Player II, so it is all that is left. And given that ll is all that remains, Du strongly dominates Bu. So the iterative deletion of weakly dominated strategies leaves us with $\langle Du, ll \rangle$ as the unique solution of the game.

Alternatively, we can think the players reason as follows. (As Stalnaker notes, this is an instance of the forward induction reasoning we discussed earlier.) Playing Bd is obviously irrational for Player I, since its maximum return is -1, and playing Du or Dd guarantees getting at least 0. So any rational player who plays B must be playing Bu. That is, if Player I is rational and plays B, she will play Bu. Moreover, this is common knowledge among the players. So if Player I plays B, she will recognise that Player II will know that she is playing Bu. And if Player II knows Player I is playing u after burning, she will play l, since that returns her 1 rather than 0. So Player I knows that playing B leads to the 2,1 state; i.e., it returns her 2. But now it is irrational to play Dd, since that gets at most 1, and playing B leads to a return of 2. So Player I will play Bu or Du. And since the reasoning that leads to this is common knowledge, Player II must play ll, since playing l is her best response to a play of u, and she knows Player I will play u. But if Player II is going to play ll, Player I doesn't need to burn; she can just play Du.

There is a crucial conditional in the middle of that argument; let's isolate it.

- If Player I is rational and plays B, she will play Bu.

But that's not what is needed to make the argument work. What Player I needs, and in fact needs to be common knowledge, is the following subjunctive.

- If Player I is rational then, if she were to play B, she would play Bu.

But in fact there's no reason to believe that. After all, if the forward induction argument is right, then it is *irrational* to burn the money. So if Player I were to play B, i.e., burn the money, then she would be doing something irrational. And the fact that Player I is actually rational is consistent with the counterfactual that if she were to do one irrational thing, it would be because she is following an irrational strategy. Note that on the assumption that Du is optimal, then if Player I finds herself having played B, it isn't clear that playing d is irrational; it isn't like it is dominated by u.

So it isn't true that if Player I were to burn the utils, Player II would know that she is playing Bu, and react accordingly by playing l. And if that's not true, then there's no reason to think that Bu is better than Dd. And if there's no reason to think that, there's no reason to be confident Player 2 will play ll.

Stalnaker goes further and provides a positive model where the players start with a common belief in rationality, and in what each other will do, but in which the players play Bu and rl. I'll leave it as an exercise to work out what counterfactuals the agents have to believe to make this rational, but suffice to say that it could be a perfectly rational solution.

There is another, relatively simple, equilibrium to the game. Player I will play Dd, and Player II will play rr. Player II is certain that Player I will play d, and will keep that belief even if Player I irrationally burns 2 utils to start with. Given that certainty, it is rational for Player I to play Dd, so Player II's belief is consistent with believing Player I to be rational. And since Player I is playing Dd, playing rr is perfectly rational for Player II; indeed it results in her best possible outcome. Moreover, given that Player II is playing rr, it makes sense for Player I to play d even if they were, irrationally, to burn the utils. So even though playing Bd would be irrational, since it is strictly dominated, if they were to (irrationally) play D, they *should* follow that by playing d. There's an important point about the scope of the claim that playing Bd is irrational. It *doesn't* imply that if the player were to irrationally play B, it would be irrational to follow with a play of d. If Player I was convinced that Player II was playing rr, then it would be rational to follow B with d.

12.3 Iterated Prisoners' Dilemma

Let's start with a slightly relabeled version of Game 1, where we use C and D for cooperate and defect.

Game 1	c	d
C	3, 3	0, 5
D	5, 0	1, 1

We'll call the players I and II, and use uppercase letters for Player I's strategies, and lower case letters for Player II's strategies. We'll assume that each player knows that the other players are perfectly rational in Stalnaker's sense.

There are a couple of arguments that the players must end up at the equilibrium where every player defects on every move. One of these is an argument by backwards induction.

At the last move of the game, defecting dominates cooperating. Both players know this. At the second-last move, you might have thought antecedently that there was some benefit to cooperating. After all, it might induce cooperation in the other player. But the only benefit one could get from cooperating was cooperation at the next (i.e., last) move. And no rational player will cooperate on the last move. So, if you're playing with a rational player, there's no benefit to cooperating on the second-last move. But if that's right, and everyone knows it, then there is no benefit to cooperating on the third-last move. The only benefit would be if that would induce cooperation, and it couldn't, since we just proved that any rational player would defect on the second-last move. And so on for any finite length game that you like.

Alternatively, we could look at the game from a strategic perspective. As we've seen in games like Game 43, sometimes there are Nash equilibria that don't appear in backwards induction reasoning. But this isn't the case in finitely iterated Prisoners' Dilemma. The only Nash equilibrium is that both players defect in every circumstance.

This is a little tricky to check since the strategic form of iterated Prisoners' Dilemma gets very complex very quickly. Let's just consider the three round version of the game. Already each player has 128 strategies to choose from. The choice of a strategy involves making 7 distinct choices:

1. What to do in the first round.
2. What to do in the second round if the other player cooperates in the first round.
3. What to do in the second round if the other player defects in the first round.
4. What to do in the third round if the other player cooperates in each of the first two rounds.
5. What to do in the third round if the other player cooperates in the first round then defects in the second round.
6. What to do in the third round if the other player defects in the first round then cooperates in the second round.
7. What to do in the third round is the other player defects in each fo the first two rounds.

Since these 7 choices are distinct, and the player has 2 choices at each point, there are $2^7 = 128$ possible strategies. So the strategic form of the table involves $128 \times 128 = 16384$ cells. Needless to say, we *won't* be putting that table here. (Though it isn't too hard to get a computer to draw the table for you. The four round game, where each player has 2^{15} choices, and there are over one billion cells in the decision table, requires more computing power!)

We'll write a strategy as $\langle x_1 x_2 x_3 x_4 x_5 x_6 x_7 \rangle$, where x_i is 0 if the player's answer to the i'th question above is to cooperate, and 1 if it is to defect. So $\langle 0000000 \rangle$ is the strategy of always cooperating, $\langle 1111111 \rangle$ is the strategy of always defecting, ⟨0010101⟩ is 'tit-for-tat', the strategy of cooperating on the first move, then copying the other player's previous move, and so on.

Let $\langle x_1x_2x_3x_4x_5x_6x_7\rangle$ be any strategy that doesn't always involve defection on the final round, i.e., a strategy where $x_4 + x_5 + x_6 + x_7 < 4$. It is easy enough to verify that such a strategy is weakly dominated by $\langle x_1x_2x_3 1111\rangle$. In some plays of the game, the defection on the final round leads to getting a better outcome. In other plays of the game, $\langle x_1x_2x_3x_4x_5x_6x_7\rangle$ and $\langle x_1x_2x_3 1111\rangle$ have the same play. For instance, ⟨0001000⟩ will do just as well as ⟨0001111⟩ if the opponent plays any strategy of the form $\langle 000x_4x_5x_6x_7\rangle$, but it can never do better than ⟨0001111⟩. So if each player is perfectly rational, we can assume their strategy ends with 1111.

That cuts each player's choices down to 8. Let's do that table. We'll use C and c rather than 0, and D and d rather than 1, so it is easier to distinguish the two players' strategies. When we label the table, we'll leave off the trailing Ds and d's, since we assume players are defecting on the last round. We'll also leave off the 3 units the players each get in the last round. (In other words, this will look a lot like the table for a *two* round iterated Prisoners' Dilemma, but it is crucial that it is actually a three-round game.)

Game 44	ccc	ccd	cdc	cdd	dcc	dcd	ddc	ddd
CCC	6, 6	6, 6	3, 8	3, 8	3, 8	3, 8	0, 10	0, 10
CCD	6, 6	6, 6	3, 8	3, 8	5, 5	5, 5	1, 6	1, 6
CDC	6, 6	8, 3	4, 4	4, 4	3, 8	3, 8	0, 10	0, 10
CDD	6, 6	8, 3	4, 4	4, 4	5, 5	5, 5	1, 6	1, 6
DCC	6, 6	5, 5	8, 3	5, 5	4, 4	1, 6	4, 4	1, 6
DCD	6, 6	5, 5	8, 3	5, 5	6, 1	2, 2	6, 1	2, 2
DDC	6, 6	6, 1	10, 0	6, 1	4, 4	1, 6	4, 4	1, 6
DDD	6, 6	6, 1	10, 0	6, 1	6, 1	2, 2	6, 1	2, 2

In this game CCC is *strongly* dominated by CDD, and DCC is strongly dominated by DDD. Similarly ccc and dcc are strongly dominated. So let's delete them.

Game 44$'$	ccd	cdc	cdd	dcd	ddc	ddd
CCD	6, 6	3, 8	3, 8	5, 5	1, 6	1, 6
CDC	8, 3	4, 4	4, 4	3, 8	0, 10	0, 10
CDD	8, 3	4, 4	4, 4	5, 5	1, 6	1, 6
DCD	5, 5	8, 3	5, 5	2, 2	6, 1	2, 2
DDC	6, 1	10, 0	6, 1	1, 6	4, 4	1, 6
DDD	6, 1	10, 0	6, 1	2, 2	6, 1	2, 2

Note that each of these is a best response. In particular,

- CCD is a best response to dcd.
- CDC is a best response to ccd.
- CDD is a best response to ccd and dcd.
- DCD is a best response to ddc.
- DDC is a best response to cdc and cdd.
- DDD is a best response to cdc, cdd, ddc and ddd.

Now the only Nash equilibrium of that table is the bottom-right corner. But there are plenty of strategies that, in a strategic version of the game, would be consistent with common belief in perfect rationality. The players could, for instance, play CDC and cdc, thinking that the other players are playing ccd and CCD respectively. Those beliefs would be false, but they wouldn't be signs that the other players are irrational, or that they are thinking the other players are irrational.

But you might suspect that the strategic form of the game and the extensive form are crucially different. The very fact that there are Nash equilibria that are not subgame perfect equilibria suggests that there are tighter constraints on what can be played in an extensive game consistent with rationality and belief in rationality. Stalnaker argues, however, that this isn't right. In particular, any strategy that is (given the right beliefs) perfectly rational in the strategic form of the game is also (given the right beliefs and belief updating dispositions) perfectly rational in the extensive form. We'll illustrate this by working more slowly through the argument that the game play $\langle CDC, cdc\rangle$ is consistent with common belief in perfect rationality.

The first thing you might worry about is that it isn't clear that CDC is perfectly rational, since it is looks to be weakly dominated by CDD, and perfect rationality is inconsistent with playing weakly dominated strategies. But in fact CDC isnt weakly dominated by CDD. It's true that on the six columns represented here, CDC never does better than CDD. But remember that each of these strategies is short for a *three*-round strategy; they are really short for $CDCDDDD$ and $CDDDDDD$. And $CDCDDDD$ is not weakly dominated by $CDDDDDD$; it does better, for example, against $dcccddd$. That is the strategy is defecting the first round, cooperating the second, then defecting on the third round unless the other player has cooperated on each of the first two rounds. Since CDC does cooperate each of the first two rounds, it gets the advantage of defecting against a cooperator on the final round, and ends up with 8 points, whereas CDD merely ends up with 6.

But why would we think Player II might play $dcccddd$? After all, it is itself a weakly dominated strategy, and perfectly rational beings (like Player II) don't play weakly dominated strategies. But we're not actually thinking Player II *will* play that. Remember, Player I's assumption is that Player II will play *ccdddddd*, which is not a weakly dominated strategy. (Indeed, it is tit-for-tat-minus-one, which is a well-known strategy.) What Player I also thinks is that if she's wrong about what Player II will play, then it is possible that Player II is not actually perfectly rational. That's consistent with believing Player II is actually perfectly rational. The assumption of common belief in perfect rationality is not sufficient for the assumption that one should believe that the other player is perfectly rational *no matter what surprises happen*. Indeed, that assumption is barely coherent; some assumptions are inconsistent with perfect rationality.

One might be tempted by a weaker assumption. Perhaps a player should hold on to the belief that the other player is perfectly rational unless they get evidence that is *inconsistent* with that belief. But it isn't clear what could motivate that. In general, when we are surprised, we have to give up something that isn't required by the surprise. If we antecedently believe $p \wedge q$, and learn $\neg(p \wedge q)$, then what we've learned is inconsistent with neither p nor q, but we must give up one of those beliefs. Similarly here, it seems we must be prepared to give up a belief in the perfect rationality of the other player in some circumstances when that is not entailed by our surprise.

What happens in the extensive form version of the game? Well, each player cooperates, thinking this will induce cooperation in the other, then each player is surprised by a defection at round two, then at the last round they both defect because it is a one-shot Prisoners' Dilemma. Nothing seems irrational there. We might wonder why neither defected at round one. Well, if they believed that the other player was playing tit-for-tat-minus-one, then it is better to cooperate at round one (and collect 8 points over the first two rounds) than to defect (and collect at most 6). And playing tit-for-tat-minus-one is rational if the other person is going to defect on the first and third rounds, and play tit-for-tat on round two. So as long as Player I thinks that Player II thinks that Player I thinks that she is going to defect on the first and third rounds, and play tit-for-tat on round two, then her cooperation at round one is rational, and consistent with believing that the other player is rational.

But note that we had to attribute an odd belief to Player II. We had to assume that Player II is *wrong* about what Player I will play. It turns out, at least for Prisoners' Dilemma, that this is crucial. If both players are certain that both players are perfectly rational, and have no false beliefs, then the only strategy that can be rationalised is permanent defection. The proof (which I'm leaving out) is in Stalnaker's "Knowledge, Belief and Counterfactual Reasoning in Games".

This *isn't* because the no false beliefs principle suffices for backwards induction reasoning in general. In Game 42, we can have a model of the game where both players are perfectly rational, and have correct beliefs about what the other player will do, and both those things are common belief, and yet the backwards induction solution is not played. In that game A believes, truly, that B will play d, and B believes, truly, that A will play A_1D_2. And each of these moves is optimal given (true!) beliefs about the other player's strategy.

But Prisoners' Dilemma is special. It isn't that in a game played between players who know each other to have no false beliefs and be perfectly rational that we must end up at the bottom-right corner of the strategic table. But we must end up in a game where every player defects every time. It could be that Player I thinks Player II is playing either dcd or ddd, with ddd being much more, and on that basis she decides to play dcd. And Player II could have the converse beliefs. So we'll end up with the play being $\langle DCD, dcd\rangle$. But of course that means each player will defect on the first two rounds, and then again on the last round since they are perfectly rational. In such a case the requirement that each player have no false beliefs won't even be enough to get us to Nash equilibrium since $\langle DCD, dcd\rangle$ is not a Nash equilibrium. But it is enough to get permanent defection.

There have been two large themes running through this section, and I'll close by separating them out.

- To justify backward induction reasoning, we need very strong assumptions. In particular, each player needs to not just believe in common knowledge of rationality, but be certain that no matter what move were to be made, there would still be common knowledge of rationality. This assumption is very implausible in anything like a real world situation, even in situations where there is actually common knowledge of rationality.
- To justify the 'always defect' solution to Iterated Prisoners' Dilemma, we don't need to offer a full justification of backward induction reasoning. Some weaker assumptions, which may be more plausible in some real-world settings, will suffice.

On those notes, we'll leave game theory, and move to looking at the theory of group decision making.

Chapter 13

Group Decisions

So far, we've been looking at the way that an individual may make a decision. In practice, we are just as often concerned with group decisions as with individual decisions. These range from relatively trivial concerns (e.g. Which movie shall we see tonight?) to some of the most important decisions we collectively make (e.g. Who shall be the next President?). So methods for grouping individual judgments into a group decision seem important.

Unfortunately, it turns out that there are several challenges facing any attempt to merge preferences into a single decision. In this chapter, we'll look at various approaches that different groups take to form decisions, and how these different methods may lead to different results. The different methods have different strengths and, importantly, different weaknesses. We might hope that there would be a method with none of these weaknesses. Unfortunately, this turns out to be impossible.

One of the most important results in modern decision theory is the Arrow Impossibility Theorem, named after the economist Kenneth Arrow who discovered it. The Arrow Impossibility Theorem says that there is no method for making group decisions that satisfies a certain, relatively small, list of desiderata. The next chapter will set out the theorem, and explore a little what those constraints are.

Finally, we'll look a bit at real world voting systems, and their different strengths and weaknesses. Different democracies use quite different voting systems to determine the winner of an election. (Indeed, within the United States there is an interesting range of systems used.) And some theorists have promoted the use of yet other systems than are currently used. Choosing a voting system is not quite like choosing a method for making a group decision. For the next two chapters, when we're looking at ways to aggregate individual preferences into a group decision, we'll assume that we have clear access to the preferences of individual agents. A voting system is not meant to tally preferences into a decision, it is meant to tally votes. And voters may have reasons (some induced by the system itself) for voting in ways other than their preferences. For instance, many voters in American presidential elections vote for their preferred candidate of the two major candidates, rather than 'waste' their vote on a third party candidate.

For now we'll put those problems to one side, and assume that members of the group express themselves honestly when voting. Still, it turns out there are complications that arise for even relatively simple decisions.

13.1 Making a Decision

Seven friends, who we'll imaginatively name $F_1, F_2, ..., F_7$ are trying to decide which restaurant to go to. They have four options, which we'll also imaginatively name R_1, R_2, R_3, R_4. The first thing they do is ask which restaurant each person prefers. The results are as follows.

- F_1, F_2 and F_3 all vote for R_1, so it gets 3 votes
- F_4 and F_5 both vote for R_2, so it gets 2 votes
- F_6 votes for R_3, so it gets 1 vote
- F_7 votes for R_4, so it gets 1 vote

It looks like R_1 should be the choice then. It, after all, has the most votes. It has a 'plurality' of the votes - that is, it has the most votes. In most American elections, the candidate with a plurality wins. This is sometimes known as plurality voting, or (for unclear reasons) first-past-the-post or winner-take-all. The obvious advantage of such a system is that it is easy enough to implement.

But it isn't clear that it is the ideal system to use. Only 3 of the 7 friends wanted to go to R_1. Possibly the other friends are all strongly opposed to this particular restaurant. It seems unhappy to choose a restaurant that a majority is strongly opposed to, especially if this is avoidable.

So the second thing the friends do is hold a 'runoff' election. This is the method used for voting in some U.S. states (most prominently in Georgia and Louisiana) and many European countries. The idea is that if no candidate (or in this case no restaurant) gets a majority of the vote, then there is a second vote, held just between the top two vote getters. (These runoffs are common in Europe, but rarer in the United States. In 2008, a runoff election determined the election of a United States Senator from Georgia.) Since R_1 and R_2 were the top vote getters, the choice will just be between those two. When this vote is held the results are as follows.

- F_1, F_2 and F_3 all vote for R_1, so it gets 3 votes

- F_4, F_5, F_6 and F_7 all vote for R_2, so it gets 4 votes

This is sometimes called 'runoff' voting, for the natural reason that there is a runoff. Now we've at least arrived at a result that the majority may not have as their first choice, but which a majority are at least happy to vote for.

But both of these voting systems seem to put a lot of weight on the various friends' first preferences, and less weight on how they rank options that aren't optimal for them. There are a couple of notable systems that allow for these later preferences to count. For instance, here is how the polls in American college sports work. A number of voters rank the best teams from 1 to n, for some salient n in the relevant sport. Each team then gets a number of points per ballot, depending on where it is ranked, with n points for being ranked first, $n-1$ points for being ranked second, $n-2$ points for being ranked third, and so on down to 1 point for being ranked n'th. The teams' overall ranking is then determined by who has the most points.

In the college sport polls, the voters don't rank every team, only the top n, but we can imagine doing just that. So let's have each of our friends rank the restaurants in order, and we'll give 4 points to each restaurant that is ranked first, 3 to each second place, etc. The points that each friend awards are given by the following table.

	$\mathbf{F_1}$	$\mathbf{F_2}$	$\mathbf{F_3}$	$\mathbf{F_4}$	$\mathbf{F_5}$	$\mathbf{F_6}$	$\mathbf{F_7}$	**Total**
$\mathbf{R_1}$	4	4	4	1	1	1	1	16
$\mathbf{R_2}$	1	3	3	4	4	2	2	19
$\mathbf{R_3}$	3	2	2	3	3	4	3	20
$\mathbf{R_4}$	2	1	1	2	2	3	4	15

Now we have yet a different choice. By this method, R_3 comes out as the best option. This voting method is sometimes called the Borda count. The nice advantage of it is that it lets all preferences, not just first preferences, count. Note that previously we didn't look at all at the preferences of the first three friends, beside noting that R_1 is their first choice. Note also that R_3 is no one's least favourite option, and is many people's second best choice. These seem to make it a decent choice for the group, and it is these facts that the Borda count is picking up on.

But there is something odd about the Borda count. Sometimes when we prefer one restaurant to another, we prefer it by just a little. Other times, the first is exactly what we want, and the second is, by our lights, terrible. The Borda count tries to approximately measure this - if X strongly prefers A to B, then often there will be many choices between A and B, so A will get many more points on X's ballot. But this is not necessary. It is possible to have a strong preference for *A* over *B* without there being any live option that is 'between' them. In any case, why try to come up with some proxy for strength of preference when we can measure it directly?

That's what happens if we use 'range voting'. Under this method, we get each voter to give each option a score, say a number between 0 and 10, and then add up all the scores. This is, approximately, what's used in various sporting competitions that involve judges, such as gymnastics or diving. In those sports there is often some provision for eliminating the extreme scores, but we won't be borrowing that feature of the system. Instead, we'll just get each friend to give each restaurant a score out of 10, and add up the scores. Here is how the numbers fall out.

	F_1	**F_2**	**F_3**	**F_4**	**F_5**	**F_6**	**F_7**	**Total**
R_1	10	10	10	5	5	5	0	45
R_2	7	9	9	10	10	7	1	53
R_3	9	8	8	9	9	10	2	55
R_4	8	7	7	8	8	9	10	57

Now R_4 is the choice! But note that the friends' individual preferences have not changed throughout. The way each friend would have voted in the previous 'elections' is entirely determined by their scores as given in this table. But using four different methods for aggregating preferences, we ended up with four different decisions for where to go for dinner.

I've been assuming so far that the friends are accurately expressing their opinions. If the votes came in just like this though, some of them might wonder whether this is really the case. After all, F_7 seems to have had an outsized effect on the overall result here. We'll come back to this when looking at options for voting systems.

13.2 Desiderata for Preference Aggregation Mechanisms

None of the four methods we used so far are obviously crazy. But they lead to four different results. Which of these, if any, is the correct result? Put another way, what is the ideal method for aggregating preferences? One natural way to answer this question is to think about some desirable features of aggregation methods. We'll then look at which systems have the most such features, or ideally have all of them.

One feature we'd like is that each option has a chance of being chosen. It would be a very bad preference aggregation method that didn't give any possibility to, say, R_3 being chosen.

More strongly, it would be bad if the aggregation method chose an option X when there was another option Y that everyone preferred to X. Using some terminology from the game theory notes, we can express this constraint by saying our method should never choose a Pareto inferior option. Call this the **Pareto condition**.

We might try for an even stronger constraint. Some of the time, not always but some of the time, there will be an option C such than a majority of voters prefers C to X, for every alternative X. That is, in a two-way match-up between C and any other option X, C will get more votes. Such an option is sometimes called a Condorcet option, after Marie Jean Antoine Nicolas Caritat, the Marquis de Condorcet, who discussed such options. The **Condorcet condition** on aggregation methods is that a Condorcet option always comes first, if such an option exists.

Moving away from these comparative norms, we might also want our preference aggregation system to be fair to everyone. A method that said F_2 is the dictator, and F_2's preferences are the group's preferences, would deliver a clear answer, but does not seem to be particularly fair to the group. There should be **no dictators**; for any person, it is possible that the group's decision does not match up with their preference.

More generally than that, we might restrict attention to preference aggregation systems that don't pay attention to *who* has various preferences, just to *what* preferences people have. Here's one way of stating this formally. Assume that two members of the group, v_1 and v_2, swap preferences, so v_1's new preference ordering is v_2's old preference ordering and vice versa. This shouldn't change what the group's decision is, since from a group level, nothing has changed. Call this the **symmetry** condition.

Finally, we might want to impose a condition that we said is a condition we imposed on independent agents: the **irrelevance of independent alternatives**. If the group would choose A when the options are A and B, then they wouldn't choose B out of any larger set of options that also include A. More generally, adding options can change the group's choice, but only to one of the new options.

Assessing Plurality Voting

It is perhaps a little disturbing to think how few of those conditions are met by plurality voting, which is how Presidents of the USA are elected. Plurality voting clearly satisfies the **Pareto condition**. If everyone prefers A to B, then B will get no votes, and so won't win. So far so good. And since any one person might be the only person who votes for their preferred candidate, and since other candidates might get more than one vote, no one person can dictate who wins. So it satisfies **no dictators**. Finally, since the system only looks at votes, and not at who cast them, it satisfies **symmetry**.

But it does not satisfy the **Condorcet condition**. Consider an election with three candidates. A gets 40% of the vote, B gets 35% of the vote, and C gets 25% of the vote. A wins, and C doesn't even finish second. But assume also that everyone who didn't vote for C has her as their second preference after either A or B. Something like this may happen if, for instance, C is an independent moderate, and A and B are doctrinaire candidates from the major parties. Then 60% prefer C to A, and 65% prefer C to B. So C is a Condorcet candidate, yet is not elected.

A similar example shows that the system does not satisfy the **irrelevance of independent alternatives** condition. If B was not running, then presumably A would still have 40% of the vote, while C would have 60% of the vote, and would win. One thing you might want to think about is how many elections in recent times would have had the outcome changed by eliminating (or adding) unsuccessful candidates in this way.

13.3 Ranking Functions

In the rest of this chapter, we'll focus on setting out and proving Arrow's Theorem. In the next chapter, we'll use Arrow's Theorem to derive some conclusions about the design of voting systems.

The theorem is a mathematical result, and needs careful setup. We'll assume that each agent has a **complete** and **transitive** preference ordering over the options. If we say $A >_V B$ means that V prefers A to B, that $A =_V B$ means that V is indifferent between A and B, and that $A \geq_V B$ means that $A >_V B \vee A =_V B$, then these constraints can be expressed as follows.

Completeness For any voter V and options A, B, either $A \geq_V B$ or $B \geq_V A$

Transitivity For any voter V and options A, B, the following three conditions hold:

- If $A >_V B$ and $B >_V C$ then $A >_V C$
- If $A =_V B$ and $B =_V C$ then $A =_V C$
- If $A \geq_V B$ and $B \geq_V C$ then $A \geq_V C$

More generally, we assume the **substitutivity of indifferent options**. That is, if $A =_V B$, then whatever is true of the agent's attitude towards A is also true of the agent's attitude towards B. In particular, whatever comparison holds in the agent's mind between A and C holds between B and C. (The last two bullet points under transitivity follow from this principle about indifference and the earlier bullet point.)

The effect of these assumptions is that we can represent the agent's preferences by lining up the options from best to worst, with the possibility that we'll have to put two options in one 'spot' to represent the fact that the agent values each of them equally.

A **ranking function** is a function from the preference orderings of the agent to a new preference ordering, which we'll call the preference ordering of the group. We'll use the subscript $_G$ to note that it is the group's ordering we are designing. We'll also assume that the group's preference ordering is complete and transitive.

There are any number ranking functions that don't look at all like the *group's* preferences in any way. For instance, if the function is meant to work out the results of an election, we could consider the function that takes any input whatsoever, and returns a ranking that simply lists by age, with the oldest first, the second oldest second, etc. This doesn't seem like it is the group's preferences in any way. Whatever any member of the group thinks, the oldest candidate wins. What Arrow called the citizen sovereignty condition is that for any possible ranking, it should be possible to have the group end up with that ranking.

The citizen sovereignty follows from another constraint we might put on ranking functions. If everyone in the group prefers A to B, then $A >_G B$, i.e. the group prefers A to B. Earlier, we called this the **Pareto** constraint. It is sometimes called the **unanimity** constraint, but we'll stick with calling it the Pareto condition.

One way to satisfy the Pareto constraint is to pick a particular person, and make them dictator. That is, the function 'selects' a person V, and says that $A >_G B$ if and only if $A >_V B$. If everyone prefers A to B, then V will, so this is consistent with the Pareto constraint. But it also doesn't seem like a way of constructing the group's preferences. So let's say that we'd like a non-dictatorial ranking function.

The last constraint is one we discussed in the previous chapter: the **independence of irrelevant alternatives**. Formally, this means that whether $A >_G B$ is true depends only on how the voters rank A and B. So changing how the voters rank, say B and C, doesn't change what the group says about the A, B comparison.

It's sometimes thought that it would be a very good thing if the voting system respected this constraint. Let's say that you believe that if Ralph Nader had not been a candidate in the 2000 U.S. Presidential election, then Al Gore, not George Bush, would have won the election. Then you might think it is a little odd that whether Gore or Bush wins depends on who else is in the election, and not on the voters' preferences between Gore and Bush. This is a special case of the independence of irrelevant alternatives - you think that the voting system should end up with the result that it would have come up with had there been just those two candidates. If we generalise this motivation a lot, we get the conclusion that third possibilities should be irrelevant.

Unfortunately, we've now got ourselves into an impossible situation. Arrow's theorem says that any ranking function that satisfies the Pareto and independence of irrelevant alternatives constraints, has a dictator in any case where the number of alternatives is greater than 2. When there are only 2 choices, majority rule satisfies all the constraints. But nothing, other than dictatorship, works in the general case.

13.4 Cyclic Preferences

We can see why three option cases are a problem by considering one very simple example. Say there are three voters, V_1, V_2, V_3 and three choices A, B, C. The agent's rankings are given in the table below. (The column under each voter lists the choices from their first preference, on top, to their least favourite option, on the bottom.)

V$_1$	**V$_2$**	**V$_3$**
A	B	C
B	C	A
C	A	B

If we just look at the A/B comparison, A looks pretty good. After all, 2 out of 3 voters prefer A to B. But if we look at the B/C comparison, B looks pretty good. After all, 2 out of 3 voters prefer B to C. So perhaps we should say A is best, B second best and C worst. But wait! If we just look at the C/A comparison, C looks pretty good. After all, 2 out of 3 voters prefer C to A.

It might seem like one natural response here is to say that the three options should be tied. The group preference ranking should just be that $A =_G B =_G = C$. But note what happens if we say that and accept independence of irrelevant alternatives. If we eliminate option C, then we shouldn't change the group's ranking of A and B. That's what independence of irrelevant alternatives says. So now we'll be left with the following rankings.

V$_1$	**V$_2$**	**V$_3$**
A	B	A
B	A	B

By independence of irrelevant alternatives, we should still have $A =_G B$. But 2 out of 3 voters wanted A over B. The one voter who preferred B to A is making it that the group ranks them equally. That's a long way from making them a dictator, but it's our first sign that our constraints give excessive power to one voter. One other thing the case shows is that we can't have the following three conditions on our ranking function.

- If there are just two choices, then the majority choice is preferred by the group.
- If there are three choices, and they are symmetrically arranged, as in the table above, then all choices are equally preferred.
- The ranking function satisfies independence of irrelevant alternatives.

I noted after the example that V_2 has quite a lot of power. Their preference makes it that the group doesn't prefer A to B. We might try to generalise this power. Maybe we could try for a ranking function that worked strictly by consensus. The idea would be that if everyone prefers A to B, then $A >_G B$, but if there is no consensus, then $A =_G B$. Since how the group ranks A and B only depends on how individuals rank A and B, this method easily satisfies independence of irrelevant alternatives. And there are no dictators, and the method satisfies the Pareto condition. So what's the problem?

Unfortunately, the consensus method described here violates transitivity, so doesn't even produce a group preference ordering in the formal sense we're interested in. Consider the following distribution of preferences.

$\mathbf{V_1}$	$\mathbf{V_2}$	$\mathbf{V_3}$
A	A	B
B	C	A
C	B	C

Everyone prefers A to C, so by unanimity, $A >_G C$. But there is no consensus over the A/B comparison. Two people prefer A to B, but one person prefers B to A. And there is no consensus over the B/C comparison. Two people prefer B to C, but one person prefers C to B. So if we're saying the group is indifferent between any two options over which there is no consensus, then we have to say that $A =_G B$, and $B =_G C$. By transitivity, it follows that $A =_G C$, contradicting our earlier conclusion that $A >_G C$.

This isn't going to be a formal argument, but we might already be able to see a difficulty here. Just thinking about our first case, where the preferences form a cycle suggests that the only way to have a fair ranking consistent with independence of irrelevant alternatives is to say that the group only prefers options when there is a consensus in favour of that option. But the second case shows that consensus based methods do not in general produce *rankings* of the options. So we have a problem. Arrow's Theorem shows how deep that problem goes.

13.5 Proofs of Arrow's Theorem

The proofs of Arrow's Theorem, though not particularly long, are a little tricky to follow. So we won't go through them in any detail at all. But I'll sketch one proof due to John Geanakopolos of the Cowles Foundation at Yale.[1] Geanakopolos assumes that we have a ranking function that satisfies Pareto and independence of irrelevant alternatives, and aims to show that in this function there must be a dictator.

The first thing he proves is a rather nice lemma. Assume that every voter puts some option B on either the top or the bottom of their preference ranking. Don't assume they all agree: some people hold that B is the very best option, and the rest hold that it is the worst. Geanakopolos shows that in this case the ranking function must put B either at the very top or the very bottom.

To see this, assume that it isn't true. So there are some options A and C such that $A \geq_G B$ and $B \geq_G C$. Now imagine changing each voter's preferences so that C is moved above A while B stays where it is - either on the top or the bottom of that particular voter's preferences. By Pareto, we'll now have $C >_G A$, since everyone prefers C to A. But we haven't changed how any person thinks about any comparison involving B. So by independence of irrelevant alternatives, $A \geq_G B$ and $B \geq_G C$ must still be true. By transitivity, it follows that $A \geq_G C$, contradicting our conclusion that $C >_G A$.

This is a rather odd conclusion I think. Imagine that we have four voters with the following preferences.

V$_1$	**V$_2$**	**V$_3$**	**V$_4$**
B	B	A	C
A	C	C	A
C	A	B	B

By what we've proven so far, B has to come out either best or worst in the group's rankings. But which should it be? Since half the people love B, and half hate it, it seems it should get a middling ranking. One lesson of this is that independence

[1] The proof is available at http://ideas.repec.org/p/cwl/cwldpp/1123r3.html.

of irrelevant alternatives is a very strong condition, one that we might want to question.

The next stage of Geanakopolos's proof is to consider a situation where at the start everyone thinks B is the very worst option out of some long list of options. One by one the voters change their mind, with each voter in turn coming to think that B is the best option. By the result we proved above, at every stage of the process, B must be either the worst option according to the group, or the best option. B starts off as the worst option, and by Pareto B must end up as the best option. So at one point, when one voter changes their mind, B must go from being the worst option on the group's ranking to being the best option, simply in virtue of that person changing their mind.

We won't go through the rest, but the proof continues by showing that that person has to be a dictator. Informally, the idea is to prove two things about that person, both of which are derived by repeated applications of independence of irrelevant alternatives. First, this person has to retain their power to move B from worst to first whatever the other people think of A and C. Second, since they can make B jump all options by changing their mind about B, if they move B 'halfway', say they come to have the view $A >_V B >_V C$, then B will jump (in the group's ranking) over all options that it jumps over in this voter's rankings. But that's possible (it turns out) only if the group's ranking of A and C is dependent entirely on this voter's rankings of A and C. So the voter is a dictator with respect to this pair. A further argument shows that the voter is a dictator with respect to every pair, which shows there must be a dictator.

Chapter 14

Voting Systems

The Arrow Impossibility Theorem shows that we can't have everything that we want in a voting system. In particular, we can't have a voting system that takes as inputs the preferences of each voter, and outputs a preference ordering of the group that satisfies these three constraints.

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

Any voting system either won't be a function in the sense that we're interested in for Arrow's Theorem, or will violate some of those constraints. (Or both.) But still there could be better or worse voting systems. Indeed, there are many voting systems in use around the world, and serious debate about which is best. In these notes we'll look at the pros and cons of a few different voting systems.

The discussion here will be restricted in two respects. First, we're only interested in systems for making political decisions, indeed, in systems for electing representatives to political positions. We're not interested in, for instance, the systems that a group of friends might use to choose which movie to see, or that an academic department might use to hire new faculty. Some of the constraints we'll be looking at are characteristic of elections in particular, not of choices in general.

Second, we'll be looking only at elections to fill a single position. This is a fairly substantial constraint. Many elections are to fill multiple positions. The way a lot of electoral systems work is that many candidates are elected at once, with the number of representatives each party gets being (roughly) in proportion to the number of people who vote for that party. This is how the parliament is elected in many countries around the world (including, for instance, Mexico, Germany and Spain). Perhaps more importantly, it is basically the norm for new parliaments to have such kind of multi-member constituencies. But the mathematical issues get a little complicated when we look at the mechanisms for selecting multiple candidates, and we'll restrict ourselves to looking at mechanisms for electing a single candidate.

14.1 Plurality voting

By far the most common method used in America, and throughout much of the rest of the world, is plurality voting. Every voter selects one of the candidates, and the candidates with the most votes wins. As we've already noted, this is called plurality, or first-past-the-post, voting.

Plurality voting clearly does not satisfy the independence of irrelevant alternatives condition. We can see this if we imagine that the voting distribution starts off with the table on the left, and ends with the table on the right. (The three candidates are *A*, *B* and *C*, with the numbers at the top of each column representing the percentage of voters who have the preference ordering listed below it.)

40%	35%	25%
A	*B*	*C*
B	*A*	*B*
C	*C*	*A*

40%	35%	25%
A	*B*	*B*
B	*A*	*C*
C	*C*	*A*

All that happens as we go from left-to-right is that some people who previously favoured *C* over *B*, come to favour *B* over *C*. Yet this change, which is completely independent of how anyone feels about *A*, is sufficient for *B* to go from losing the election 40-35 to winning the election 60-40.

This is how we show that a system does not satisfy independent of irrelevant alternatives - coming up with a pair of situations where no voter's opinion about the relative merits of two choices (in this case *A* and *B*) changes, but the group's ranking of those two choices changes.

One odd effect of this is that whether *B* wins the election depends not just on how voters compare *A* and *B*, but on how voters compare *B* and *C*. One of the consequences of Arrow's Theorem might be taken to be that this kind of thing is unavoidable, but it is worth stopping to reflect on just how pernicious this is to the democratic system.

Imagine that we are in the left-hand situation, and you are one of the 25% of voters who like C best, then B then A. It seems that there is a reason for you to not vote the way your preferences go; you'll have a better chance of electing a candidate you prefer if you vote, against your preferences, for B. So the voting system might encourage voters to not express their preferences adequately. This can have a snowball effect - if in one election a number of people who prefer C vote for B, at future elections other people who might have voted for C will also vote for B because they don't think enough other people share their preferences for C to make such a vote worthwhile.

Indeed, if the candidate C themselves strongly prefers B to A, but thinks a lot of people will vote for them if they run, then C might even be discouraged from running because it will lead to a worse election result. This doesn't seem like a democratically ideal situation.

Some of these consequences are inevitable consequences of a system that doesn't satisfy independence of irrelevant alternatives. And the Arrow Theorem shows that it is hard to avoid independence of irrelevant alternatives. But some of them seem like serious democratic shortcomings, the effects of which can be seen in American democracy, and especially in the extreme power the two major parties have. (Though, to be fair, a number of other electoral systems that use plurality voting do not have such strong major parties. Indeed, Canada seems to have very strong third parties despite using this system.)

One clear advantage of plurality voting should be stressed: it is quick and easy. There is little chance that voters will not understand what they have to do in order to express their preferences. (Although as Palm Beach county keeps showing us, this can happen.) And voting is, or at least should be, relatively quick. The voter just has to make one mark on a piece of paper, or press a single button, to vote. When the voter is expected to vote for dozens of offices, as is usual in America (though not elsewhere) this is a serious benefit. In several recent U.S. elections we have seen queues hours long of people waiting to vote. Were voting any slower than it actually is, these queues might have been worse.

Relatedly, it is easy to count the votes in a plurality system. You just sort all the votes into different bundles and count the size of each bundle. Some of the other systems we'll be looking at are much harder to count the votes in. Even with plurality voting, it can take weeks to count a U.S. election. If the U.S. didn't use plurality voting, this would likely be a much worse problem.

14.2 Runoff Voting

One solution to some of the problems with plurality voting is runoff voting, which is used in parts of America (notably Georgia and Louisiana) and is very common throughout Europe and South America. The idea is that there are, in general, two elections. At the first election, if one candidate has majority support, then they win. But otherwise the top two candidates go into a runoff. In the runoff, voters get to vote for one of those two candidates, and the candidate with the most votes wins.

This doesn't entirely deal with the problem of a spoiler candidate having an outsized effect on the election, but it makes such cases a little harder to produce. For instance, imagine that there are four candidates, and the arrangement of votes is as follows.

35%	30%	20%	15%
A	*B*	*C*	*D*
B	*D*	*D*	*C*
C	*C*	*B*	*B*
D	*A*	*A*	*A*

In a plurality election, *A* will win with only 35% of the vote.[1] In a runoff election, the runoff will be between *A* and *B*, and presumably *B* will win, since 65% of the voters prefer *B* to *A*. But look what happens if *D* drops out of the election, or all of *D*'s supporters decide to vote more strategically.

35%	30%	20%	15%
A	*B*	*C*	*C*
B	*C*	*B*	*B*
C	*A*	*A*	*A*

Now the runoff is between *C* and *A*, and *C* will win. *D* being a candidate means that the candidate most like *D*, namely *C*, loses a race they could have won.

In one respect this is much like what happens with plurality voting. On the other hand, it is somewhat harder to find real life cases that show this pattern of votes. That's in part because it is hard to find cases where there are (a) four serious candidates, and (b) the third and fourth candidates are so close ideologically that they eat into each other's votes and (c) the top two candidates are so close that these third and fourth candidates combined could leapfrog over each of them.

[1] This isn't actually that unusual in the overall scope of American elections. In 2008, John McCain won several crucial Republican primary elections, especially in Florida and Missouri, with under 35% of the vote. Without those wins, the Republican primary contest would have been much closer.

Theoretically, the problem about spoiler candidates might look as severe, but it is much less of a problem in practice.

The downside of runoff voting of course is that it requires people to go and vote twice. This can be a major imposition on the time and energy of the voters. More seriously from a democratic perspective, it can lead to an unrepresentative electorate. In American runoff elections, the runoff typically has a much lower turnout than the initial election, so the election comes down to the true party loyalists. In Europe, the first round often has a very low turnout, which has led on occasion to fringe candidates with a small but loyal supporter base making the final round.

14.3 Instant Runoff Voting

One approach to this problem is to do, in effect, the initial election and the runoff at the same time. In instant runoff voting, every voter lists their preference ordering over their desired candidates. In practice, that means marking '1' beside their first choice candidate, '2' beside their second choice and so on through the candidates.

When the votes are being counted, the first thing that is done is to count how many first-place votes each candidate gets. If any candidate has a majority of votes, they win. If not, the candidate with the lowest number of votes is eliminated. The vote counter then distributes each ballot for that eliminated candidate to whichever candidate receives the '2' vote on that ballot. If that leads to a candidate having a majority, that candidate wins. If not, the candidate with the lowest number of votes at this stage is eliminated, and their votes are distributed, each voter's vote going to their most preferred candidate of the remaining candidates. This continues until a candidate gets a majority of the votes.

This avoids the particular problem we discussed about runoff voting. In that case, D would have been eliminated at the first round, and D's votes would all have flowed to C. That would have moved C about B, eliminating B. Then with B's preferences, C would have won the election comfortably. But it doesn't remove all problems. In particular, it leads to an odd kind of strategic voting possibility. The following situation does arise, though rarely. Imagine the voters are split the following way.

45%	28%	27%
A	B	C
B	A	B
C	C	A

As things stand, C will be eliminated. And when C is eliminated, all of C's votes will be transferred to B, leading to B winning. Now imagine that a few of A's voters change the way they vote, voting for C instead of their preferred candidate A, so now the votes look like this.

43%	28%	27%	2%
A	*B*	*C*	*C*
B	*A*	*B*	*A*
C	*C*	*A*	*B*

Now C has more votes than B, so B will be eliminated. But B's voters have A as their second choice, so now A will get all the new votes, and A will easily win. Some theorists think that this possibility for strategic voting is a sign that instant runoff voting is flawed.

Perhaps a more serious worry is that the voting and counting system is more complicated. This slows down voting itself, though this is a problem can be partially dealt with by having more resources dedicated to making it possible to vote. The vote count is also somewhat slower. A worse consequence is that because the voter has more to do, there is more chance for the voter to make a mistake. In some jurisdictions, if the voter does not put a number down for each candidate, their vote is invalid, even if it is clear which candidate they wish to vote for. It also requires the voter to have opinions about all the candidates running, and this may include a number of frivolous candidates. But it isn't clear that this is a major problem if it does seem worthwhile to avoid the problems with plurality and runoff voting.

So far, we have looked at a number of voting systems that are in widespread use in various democracies. We'll end by looking at three voting systems that are not used for mass elections anywhere around the world, though all of them have been used in various purposes for combining the views of groups. (For instance, they have been used for elections in small groups.)

14.4 Borda Count

In a Borda Count election, each voter ranks each of the candidates, as in Instant Runoff Voting. Each candidate then receives n points for each first place vote they receive (where n is the number of candidates), $n - 1$ points for each second place vote, and so on through the last place candidate getting 1 point. The candidate with the most points wins.

One nice advantage of the Borda Count is that it eliminates the chance for the kind of strategic voting that exists in Instant Runoff Voting, or for that matter any kind of Runoff Voting. It can never make it more likely that A will win by someone changing their vote away from A. Indeed, this could only lead to A having fewer votes. This certainly seems to be reasonable.

Another advantage is that many preferences beyond first place votes count. A candidate who is every single voter's second best choice will not do very well under any voting system that gives a special weight to first preferences. But such a candidate may well be in a certain sense the best representative of the whole community.

And a third advantage is that the Borda Count includes a rough approximation of voter's strength of preference. If one voter ranks A a little above B, and another votes B many places above A, that's arguably a sign that B is a better representative of the two of them than A. Although only one of the two prefers B, one voter will be a little disappointed that B wins, while the other would be very disappointed if B lost.

These are not trivial advantages. But there are also many disadvantages which explain why no major electoral system has adopted Borda Count voting yet, despite its strong support from some theorists.

First, Borda Count is particularly complicated to implement. It is just as difficult for the voter to as in Instant Runoff Voting; in each case they have to express a complete preference ordering. But it is much harder to count, because the vote counter has to detect quite a bit of information from each ballot. Getting this information from millions of ballots is not a trivial exercise.

Second, Borda Count has a serious problem with 'clone candidates'. In plurality voting, a candidate suffers if there is another candidate much like them on the ballot. In Borda Count, a candidate can seriously gain if such a candidate is added. Consider the following situation. In a certain electorate, of say 100,000 voters, 60% of the voters are Republicans, and 40% are Democrats. But there is only one Republican, call them R, on the ballot, and there are 2 Democrats, $D1$ and $D2$ on the ballot. Moreover, $D2$ is clearly a worse candidate than $D1$, but the Democrats still prefer the Democrat to the Republican. Since the district is overwhelmingly Republican, intuitively the Republican should win. But let's work through what happens if 60,000 Republicans vote for R, then $D1$, then $D2$, and the 40,000 Democrats vote $D1$ then $D2$ then R. In that case, R will get $60,000 \times 3 + 40,000 \times 1 = 220,000$ points, $D1$ will get $60,000 \times 2 + 40,000 \times 3 = 240,000$ points, and $D2$ will get $60,000 \times 1 + 40,000 \times 2 = 140,000$ points, and $D1$ will win. Having a 'clone' on the ticket was enough to push $D1$ over the top.

On the one hand, this may look a lot like the mirror image of the 'spoiler' problem for plurality voting. But in another respect it is much worse. It is hard to get someone who is a lot ideologically like your opponent to run in order to improve your electoral chances. It is much easier to convince someone who already wants you to win to add their name to the ballot in order to improve your chances. In practice, this would either lead to an arms race between the two parties, each trying to get the most names onto the ballot, or very restrictive (and hence undemocratic) rules about who was even allowed to be on the ballot, or, most likely, both.

The third problem comes from thinking through the previous problem from the point of view of a Republican voter. If the Republican voters realise what is up, they might vote tactically for $D2$ over $D1$, putting R back on top. In a case where the electorate is as partisan as in this case, this might just work. But this means that Borda Count is just as susceptible to tactical voting as other systems; it is just that the tactical voting often occurs downticket. (There are more complicated problems, that we won't work through, about what happens if the voters mistakenly judge what is likely to happen in the election, and tactical voting backfires.)

Finally, it's worth thinking about whether the feature of Borda Count that supporters claim as its major virtue, the fact that it considers all preferences and not just first choices, is a real gain. The core idea behind Borda Count is that all preferences should count equally. So the difference between first place and second place in a voter's affections counts just as much as the difference between third and fourth. But for many elections, this isn't how the voters themselves feel. I suspect many people reading this have strong feelings about who was the best candidate in the past Presidential election. I suspect very few people had strong feelings about who was the third best versus fourth best candidate. This is hardly a coincidence; people identify with a party that is their first choice. They say, "I'm a Democrat" or "I'm a Green" or "I'm a Republican". They don't identify with their third versus fourth preference. Perhaps voting systems that give primary weight to first place preferences are genuinely reflecting the desires of the voters.

14.5 Approval Voting

In plurality voting, every voter gets to vote for one candidate, and the candidate with the most votes wins. Approval voting is similar, except that each voter is allowed to vote for as many candidates as they like. The votes are then added up, and the candidate with the most votes wins. Of course, the voter has an interest in not voting for too many candidates. If they vote for all of the candidates, this won't advantage any candidate; they may as well have voted for no candidates at all.

The voters who are best served by approval voting, at least compared to plurality voting, are those voters who wish to vote for a non-major candidate, but who also have a preference between the two major candidates. Under approval voting, they can vote for the minor candidate that they most favour, and also vote for the the major candidate who they hope will win. Of course, runoff voting (and Instant Runoff Voting) also allow these voters to express a similar preference. Indeed, the runoff systems allow the voters to express not only two preferences, but express the order in which they hold those preferences. Under approval voting, the voter only gets to vote for more than one candidate, they don't get to express any ranking of those candidates.

But arguably approval voting is easier on the voter. The voter can use a ballot that looks just like the ballot used in plurality voting. And they don't have to learn about preference flows, or Borda Counts, to understand what is going on in the voting. Currently there are many voters who vote for, or at least appear to try to vote for, multiple candidates. This is presumably inadvertent, but approval voting would let these votes be counted, which would refranchise a number of voters. Approval voting has never been used as a mass electoral tool, so it is hard to know how quick it would be to count, but presumably it would not be incredibly difficult.

One striking thing about approval voting is that it is not a function from voter preferences to group preferences. Hence it is not subject to the Arrow Impossibility Theorem. It isn't such a function because the voters have to not only rank the candidates, they have to decide where on their ranking they will 'draw the line' between candidates that they will vote for, and candidates that they will not vote for. Consider the following two sets of voters. In each case candidates are listed from first preference to last preference, with stars indicating which candidates the voters vote for.

40%	35%	25%
$*A$	$*B$	$*C$
B	A	B
C	C	A

40%	35%	25%
$*A$	$*B$	$*C$
B	A	$*B$
C	C	A

In the election on the left-hand-side, no voter takes advantage of approval voting to vote for more than one candidate. So A wins with 40% of the vote. In the election on the right-hand-side, no one's preferences change. But the 25% who prefer C also decide to vote for B. So now B has 60% of the voters voting for them, as compared to 40% for A and 25% for C, so B wins.

This means that the voting system is not a function from voter preferences to group preferences. If it were a function, fixing the group preferences would fix who wins. But in this case, without a single voter changing their preference ordering of the candidates, a different candidate won. Since the Arrow Impossibility

Theorem only applies to functions from voter preferences to group preferences, it does not apply to Approval Voting.

14.6 Range Voting

In Range Voting, every voter gives each candidate a score. Let's say that score is from 0 to 10. The name 'Range' comes from the range of options the voter has. In the vote count, the score that each candidate receives from each voter is added up, and the candidate with the most points wins.

In principle, this is a way for voters to express very detailed opinions about each of the candidates. They don't merely rank the candidates, they measure how much better each candidate is than all the other candidates. And this information is then used to form an overall ranking of the various candidates.

In practice, it isn't so clear this would be effective. Imagine that a voter V thinks that candidate A would be reasonably good, and candidate B would be merely OK, and that no other candidates have a serious chance of winning. If V was genuinely expressing their opinions, they might think that A deserves an 8 out of 10, and B deserves a 5 out of 10. But V wants A to win, since V thinks A is the better candidate. And V knows that what will make the biggest improvement in A's chances is if they score A a 10 out of 10, and B a 0 out of 10. That will give A a 10 point advantage, whereas they may only get a 3 point advantage if the voter voted sincerely.

It isn't unusual for a voter to find themselves in V's position. So we might suspect that although Range Voting will give the voters quite a lot of flexibility, and give them the chance to express detailed opinions, it isn't clear how often it would be in a voter's interests to use these options.

And Range Voting is quite complex, both from the perspective of the voter and of the vote counter. There is a lot of information to be gleaned from each ballot in Range Voting. This means the voter has to go to a lot of work to fill out the ballot, and the vote counter has to do a lot of work to process all that information. This means that Range Voting might be very slow, both in terms of voting and counting. And if voters have a tactical reason for not wanting to fill in detailed ballots, this might mean it's a lot of effort for not a lot of reward, and that we should stick to somewhat simpler vote counting methods.