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	(2, 1)	(0, 0)
Playhouse	(0, 0)	(1, 2)

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.

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: Sleeping Beauty

Sleeping Beauty is about to undergo a slightly complicated experiment. It is now Sunday night, and a fair coin is about to be tossed, though Sleeping Beauty won't see how it lands. Then she will be asked a question, and then she'll go to sleep. She'll be woken up on Monday, asked the same question, and then she'll go back to sleep, and her memory of being woken on Monday will be wiped. Then, if (and only if) the coin landed tails, she'll be woken on Tuesday, and asked the same question, and then she'll go back to sleep. Finally, she'll wake on Wednesday.

The question she'll be asked is **How probable do you think it is that the coin landed heads?** What answers should she give

1. When she is asked on Sunday?
2. When she is asked on Monday?
3. If she is asked on Tuesday?

It seems plausible to suggest that the answers to questions 2 and 3 should be the same. After all, given that Sleeping Beauty will have forgotten about the Monday waking if she wakes on Tuesday, then she won't be able to tell the difference between the Monday and Tuesday waking. So she should give the same answers on Monday and Tuesday. We'll assume that in what follows.

First, there seems to be a very good argument for answering $\frac{1}{2}$ to question 1. It's a fair coin, so it has a probability of $\frac{1}{2}$ of landing heads. And it has just been tossed, and there hasn't been any 'funny business'. So that should be the answer.

Second, there seems to be a good, if a little complicated, argument for answering $\frac{1}{3}$ to questions 2 and 3. Assume that questions 2 and 3 are in some sense the same question. And assume that Sleeping Beauty undergoes this experiment many times. Then she'll be asked the question twice as often when the coin lands tails as when it lands heads. That's because when it lands tails, she'll be asked that question twice, but only once when it lands heads. So only $\frac{1}{3}$ of the time when she's asked this question, will it be true that the coin landed heads. And plausibly, if you're going to be repeatedly asked *How probable is it that such-and-such happened*, and $\frac{1}{3}$ of the time when you're asked that question, such-and-such will have happened, then you should answer $\frac{1}{3}$ each time.

Finally, there seems to be a good argument for answering questions 1 and 2 the same way. After all, Sleeping Beauty doesn't learn anything new between the two questions. She wakes up, but she knew she was going to wake up. And she's asked the question, but she knew she was going to be asked the question. And it seems like a decent principle that if nothing happens between Sunday and Monday to give you new evidence about a proposition, the probability that you think it did happen shouldn't change.

But of course, these three arguments can't all be correct. So we have to decide which one is incorrect.

Upcoming These are just two of the puzzles we'll be looking at as the course proceeds. Some of these will be decision puzzles, like Newcomb's Problem. Some of them will be probability puzzles that are related to decision theory, like Sleeping Beauty. And some will be game puzzles. I hope the puzzles are somewhat interesting. I hope even more that we learn something from them.