Belief, Desire, and Rational Choice

Wolfgang Schwarz

April 25, 2019

Contents

1 Modelling Rational Agents

1.1 Overview

We are going to study a general model of belief, desire, and rational choice. At the heart of this model lies a certain conception of how beliefs and desires combine to produce actions.

Let's start with an example.

Example 1.1 (The Miner Problem)

Ten miners are trapped in a shaft and threatened by rising water. You don't know whether the miners are in shaft A or in shaft B. You can block the water from entering one shaft, but you can't block both. If you block the correct shaft, all ten will survive. If you block the wrong shaft, all of them will die. If you do nothing, one miner (the shortest of the ten) will die.

What should you do?

There's a sense in which the answer depends on the location of the miners. If the miners are in shaft A, it's best to block shaft A; if they are in B, you should block B. The problem is that you need to make your choice without knowing where the miners are. You can't let your choice be guided by the unknown location of the miners. The question on which we will focus is therefore not what you should do *in light of all the facts*, but what you should do *in light of your information*. In other words, we want to know what a rational agent would do in your state of uncertainty.

A similar ambiguity arises for goals or values. Arguably, it is better to let one person die than to take a risk of ten people dying. But the matter isn't trivial, and many philosophers would disagree. Suppose you are one of these philosophers: you think it would be wrong to sacrifice the shortest miner. *By your values*, it would be better to block either shaft A or shaft B.

When we ask what an agent should do in a given decision problem, we will always mean what they should do in light of whatever they believe about their

situation and of whatever goals or values they happen to have. We will also ask whether those beliefs and goals are themselves reasonable. But it is best to treat these as separate questions.

So we have three questions:

1. How should you act so as to further your goals in the light of your beliefs?
2. What should you believe?
3. What should you desire? What are rational goals or values?

These are big questions. By the end of this course, we will not have found complete and definite answers, but we will at least have clarified the questions and made some progress towards an answer.

To begin, let me introduce a standard format for thinking about decision problems.

Exercise 1.1 ⋆

In a surprise outbreak of small pox (a deadly infectious disease), a doctor recommends vaccination for an infant, knowing that around one in a million children die from the vaccination. The infant gets the vaccination and dies. Was the doctor's recommendation wrong? Or was it wrong in one sense and right in another? If so, can you explain these senses?

1.2 Decision matrices

In decision theory, decision problems are traditionally decomposed into three ingredients, called 'acts', 'states', and 'outcomes'.

The **acts** are the options between which the agent has to choose. In the Miner Problem, there are three acts: *blocking shaft A*, *blocking shaft B*, and *doing nothing*. ('Possible act' would be a better name: if, say, you decide to do nothing, then blocking shaft *A* is not an actual act; it's not something you do, but it's something you could have done.)

The **outcomes** are whatever might come about as a result of the agent's choice. In the Miner Problem, there are three relevant outcomes: all miners survive, all miners die, and all but one survive. (Again, only one of these will actually come about, the others are merely possible outcomes.)

Each of the three acts leads to one of the outcomes. But you don't know how the outcomes are associated with the acts. For example, you don't know whether

blocking shaft *A* would lead to all miners surviving or to all miners dying. It depends on where the miners are.

This dependency between acts and outcomes is captured by the states. A **state** is a possible circumstance on which the result of the agent's choice depends. In the Miner Problem, there are two relevant states: that the miners are in shaft *A*, and that the miners are in shaft *B*. (In real decision problems, there are often many more states, just as there are many more acts.)

We can now summarize the Miner Problem in a table, called a **decision matrix**:

	Miners in *A*	Miners in *B*
Block shaft *A*	all 10 live	all 10 die
Block shaft *B*	all 10 die	all 10 live
Do nothing	1 dies	1 dies

The rows in a decision matrix always represent the acts, the columns the states, and the cells the outcome of performing the relevant act in the relevant state.

Let's do another example.

Example 1.2 (The mushroom problem)

You find a mushroom. You're not sure whether it's a delicious *paddy straw* or a poisonous *death cap*. You wonder whether you should eat it.

Here the decision matrix might look as follows. Make sure you understand how to read the matrix.

	Paddy straw	Death cap
Eat	satisfied	dead
Don't eat	hungry	hungry

Sometimes the "states" are actions of other people, as in the next example.

Example 1.3 (The Prisoner Dilemma)

You and your partner have been arrested for some crime and are separately interrogated. If you both confess, you will both serve five years in prison. If one of you confesses and the other remains silent, the one who confesses is set free, the other has to serve eight years. If you both remain silent, you can only be convicted of obstruction of justice and will both serve one year.

The Prisoner Dilemma combines two decision problems: one for you and one for your partner. We could also think about a third problem which you face as a group. Let's focus on the decision you have to make. Your choice is between confessing and remaining silent. These are the acts. What are the possible outcomes? If you only care about your own prison term, the outcomes are 5 years, 8 years, 0 years, and 1 year. Which act leads to which outcome depends on whether your partner confesses or remains silent. These are the states. In matrix form:

	Partner confesses	Partner silent
Confess	5 years	0 years
Remain silent	8 years	1 year

Notice that if your goal is to minimize your prison term, then confessing leads to the better outcome no matter what your partner does.

I've assumed you only care about your own prison term. What if you also care about the fate of your partner? Then your decision problem is not adequately summarized by the above matrix, as the cells in the matrix don't say what happens to your partner.

The outcomes in a decision problem must always include everything that matters to the agent. So if you care about your partner's sentence, the matrix should look as follows.

	Partner confesses	Partner silent
Confess	you 5 years, partner 5 years	you 0 years, partner 8 years
Remain silent	you 8 years, partner 0 years	you 1 year, partner 1 years

Now confessing is no longer the obviously best choice. For example, if your goal is to minimize the combined prison term for you and your partner, then remaining silent is better no matter what your partner does.

Exercise 1.2 ⋆

Draw the decision matrix for the game *Rock, Paper, Scissors*, assuming all you care about is whether you win.

1.3 Belief, desire, and degrees

To solve a decision problem we need to know the agent's goals and beliefs. Moreover, it is usually not enough just to know *what* the agent believes and desires; we also need to know *how strong* these attitudes are.

Let's return to the mushroom problem. Suppose you like eating a delicious mushroom, and you dislike being hungry and being dead. We might therefore label the outcomes 'good' or 'bad', reflecting your desires:

	Paddy straw	Death cap
Eat	satisfied (good)	dead (bad)
Don't eat	hungry (bad)	hungry (bad)

Now it looks like eating the mushroom is the better option: not eating is guaranteed to lead to a bad outcome, while eating at least gives you a shot at a good outcome.

The problem is that you probably prefer being hungry to being dead. Both outcomes are bad, but one is much worse than the other. So we need to represent not only the **valence** of your desires – whether an outcome is something you'd like or dislike – but also their **strength**.

An obvious way to represent both valence and strength is to label the outcomes with numbers, like so:

	Paddy straw	Death cap
Eat	satisfied (+1)	dead (-100)
Don't eat	hungry (-1)	hungry (-1)

The outcome of eating a paddy straw gets a value of +1, because it's moderately desirable. The other outcomes are negative, but death (-100) is rated much worse than hunger (-1).

The numerical values assigned to outcomes are called **utilities** (or sometimes **desirabilities**). Utilities measure the relative strength and valence of desire. We will have a lot more to say on what that means in due course.

We also need to represent the strength of your beliefs. Whether you should eat the mushroom arguably depends on how confident you are that it is a Paddy straw.

Here again we will represent the valence and strength of beliefs by numbers, but this time we'll only use numbers between 0 and 1. If the agent is certain

that a given state obtains, then her degree of belief is 1; if she is certain that the state does *not* obtain, her degree of belief is 0; if she is completely undecided, her degree of belief is 1/2. These numbers are called **credences**.

In classical decision theory, we are not interested in the agent's beliefs about the acts or the outcomes, but only in her beliefs about the states. The fully labelled mushroom matrix might therefore look as follows, assuming you are fairly confident, but by no means certain, that the mushroom is a paddy straw.

	Paddy straw (0.8)	Death cap (0.2)
Eat	satisfied (+1)	dead (-100)
Don't eat	hungry (-1)	hungry (-1)

The numbers 0.8 and 0.2 in the column headings specify your degree of belief in the two states.

The idea that beliefs vary in strength has proved fruitful not just in decision theory, but also in epistemology, philosophy of science, artificial intelligence, statistics, and other areas. The keyword to look out for is **Bayesian**: if a theory or framework is called Bayesian, this usually means it involves degrees of belief. The name refers to Thomas Bayes (1701–1761), who made an important early contribution to the movement. We will look at some applications of "Bayesianism" in later chapters.

Much of the power of Bayesian models derives from the assumption that rational degrees of belief satisfy the mathematical conditions on a probability function. Among other things, this means that the credences assigned to the states in a decision problem must add up to 1. For example, if you are 80 percent (0.8) confident that the mushroom is a paddy straw, then you can't be more than 20 percent confident that the mushroom is a death cap. It would be OK to reserve some credence for further possibilities, so that the credence in the paddy straw possibility and the death cap possibility add up to less than 1. But then our decision matrix should include further columns for the other possibilities.

So rational degrees of belief have a certain formal structure. What about degrees of desire? At first glance, these don't seem have much of a structure. For example, the fact that your utility for eating a paddy straw is +1 does not seem to entail anything about your utility for eating a death cap. Nonetheless, we will see that utilities also have a rich formal structure – a structure that is entangled with the structure of belief.

We will also discuss more substantive, non-formal constraints on belief and desire. Economists often assume that rational agents are self-interested, and so

the term 'utility' is often associated with personal wealth or welfare. That's not how we will use the term. Real people don't just care about themselves, and there is nothing wrong with that.

Exercise 1.3 ⋆

Add utilities and (reasonable) credences to your decision matrix for *Rock, Paper, Scissors*.

1.4 Solving decision problems

Suppose we have drawn up a decision matrix and filled in the credences and utilities. We then have all we need to solve the decision problem – to say what the agent should do in light of her goals and beliefs.

Sometimes the task is easy because some act is best in every state. We've already seen an example in the Prisoner Dilemma, given that all you care about is minimizing your own prison term. The fully labelled matrix might look like this:

	Partner confesses (0.5)	Partner silent (0.5)
Confess	5 years (-5)	0 years (0)
Remain silent	8 years (-9)	1 year (-1)

In the lingo of decision theory, confessing **dominates** remaining silent. In general, an act A dominates an act B if A leads to an outcome with greater utility than B in every possible state. An act is **dominant** if it dominates all other acts. If there's a dominant act, it is always the best choice (by the light of the agent).

The Prisoner Dilemma is famous because it refutes the idea that good things will always come about if people only look after their own interests. If both parties in the Prisoner Dilemma only care about themselves, they end up 5 years in prison. If they had cared enough about each other, they could have gotten away with 1.

Often there is no dominant act. Recall the mushroom problem.

	Paddy straw (0.8)	Death cap (0.2)
Eat	satisfied (+1)	dead (-100)
Don't eat	hungry (-1)	hungry (-1)

It is better to eat the mushroom if it's a paddy straw, but better not to eat it if it's a death cap. So neither option is dominant.

You might say that it's best not to eat the mushroom because eating could lead to a really bad outcome, with utility -100, while not eating at worst leads to an outcome with utility -1. This is an instance of **worst-case reasoning**. The technical term is **maximin** because worst-case reasoning tells you to choose the option that *max*imizes the *min*imal utility.

People sometimes appeal to worst-case reasoning when giving health advice or policy recommendations, and it works out OK in the mushroom problem. Nonetheless, as a general decision rule, worst-case reasoning is indefensible.

Imagine you have 100 sheep who have consumed water from a contaminated well and will die unless they're given an antidote. Statistically, one in a thousand sheep die even when given the antidote. According to worst-case reasoning there is no point of giving your sheep the antidote: either way, the worst possible outcome is that all the sheep will die. In fact, if we take into account the cost of the antidote, then worst-case reasoning suggests you should not give the antidote (even if it is cheap).

Worst-case reasoning is indefensible because it doesn't take into account the likelihood of the worst case, and because it ignores what might happen if the worst case doesn't come about. A sensible decision rule should look at all possible outcomes, paying special attention to really bad and really good ones, but also taking into account their likelihood.

The standard recipe for solving decision problems therefore evaluates each act by the *weighted average* of the utility of all possible outcomes, weighted by the likelihood of the relevant state, as given by the agent's credence.

Let's first recall how simple averages are computed. If we have n numbers $x_1, x_2, \ldots, x_n$, then the average of the numbers is

$$\frac{x_1 + x_2 + \ldots + x_n}{n} = \tfrac{1}{n} \cdot x_1 + \tfrac{1}{n} \cdot x_2 + \ldots + \tfrac{1}{n} \cdot x_n.$$

('$\cdot$' stands for multiplication.) Here each number is given the same weight, $1/n$. In a weighted average, the weights can be different for different numbers.

Concretely, to compute the weighted average of the utility that might result from eating the mushroom, we multiply the utility of each possible outcome (+1 and -100) by your credence in the corresponding state, and then add up these products. The result is called the **expected utility** of eating the mushroom.

$$EU(\text{Eat}) = 0.8 \cdot (+1) + 0.2 \cdot (-100) = -19.2.$$

In general, suppose an act A leads to outcomes $O_1, \ldots, O_n$ respectively in states $S_1, \ldots, S_n$. Let '$\mathrm{Cr}(S_1)$' denote the agent's degree of belief (or credence) in S_1, '$\mathrm{Cr}(S_2)$' her credence in S_2, etc. Let '$U(O_1)$' denote the utility of O_1 for the agent, '$U(O_2)$' the utility of O_2, etc. Then the expected utility of A is defined as

$$EU(A) = \mathrm{Cr}(S_1) \cdot U(O_1) + \ldots + \mathrm{Cr}(S_n) \cdot U(O_n).$$

You'll often see this abbreviated using the 'sum' symbol $\sum$:

$$EU(A) = \sum_{i=1}^{n} \mathrm{Cr}(S_i) \cdot U(O_i).$$

It means the same thing.

Note that the expected utility of eating the mushroom is -19.2 even though the most likely outcome has positive utility. A really bad outcome can seriously push down an act's expected utility even if the outcome is quite unlikely.

Let's calculate the expected utility of not eating the mushroom:

$$EU(\text{Not Eat}) = 0.8 \cdot -1 + 0.2 \cdot -1 = -1.$$

No surprise here. If all the numbers $x_1, \ldots, x_n$ are the same, their weighted average will again be that number.

Now we can state one of the central assumptions of our model:

The MEU Principle

Rational agents maximize expected utility.

That is, when faced with a decision problem, rational agents choose an option with greatest expected utility.

Exercise 1.4 ⋆

Assign utilities to the outcomes in the Prisoner Dilemma, assign credences to the states, and compute the expected utility of the two acts.

Exercise 1.5 ⋆

Assign utilities to the outcomes in the Miner Problem, assign credences to the states, and compute the expected utility of the three acts.

Exercise 1.6 ⋆⋆

Explain why the following decision rule is not generally reasonable: *Identify the most likely state; then choose an act which maximizes utility in that state.*

Exercise 1.7 ⋆⋆⋆

Show that if there is a dominant act, then it maximizes expected utility.

Exercise 1.8 ⋆⋆⋆

When applying dominance reasoning or the MEU Principle, it is important that the decision matrix is set up correctly.

A student wants to pass an exam and wonders whether she ought to study. She draws up the following matrix.

	Will Pass (0.5)	Won't Pass (0.5)
Study	Pass & No Fun (1)	Fail & No Fun (-8)
Don't Study	Pass & Fun (5)	Fail & Fun (-2)

She finds that not studying is the dominant option.

The student has correctly identified the acts and the outcomes in her decision problem, but the states are wrong. In an adequate decision matrix, the states must be independent of the acts: whether a given state obtains should not be affected by which act the student chooses.

Can you draw an adequate decision matrix for the student's decision problem?

Exercise 1.9 (Pascal's Wager) ⋆⋆

The first recorded use of the MEU Principle outside gambling dates back to 1653, when Blaise Pascal presented the following argument for leading a pious life. (I paraphrase.)

An impious life is more pleasant and convenient than a pious life. But if God exists, then a pious life is rewarded by salvation while an impious life is punished by eternal damnation. Thus it is rational to lead a pious life even if one gives quite low credence to the existence of God.

Draw the matrix for the decision problem as Pascal conceives it and verify that a pious life has greater expected utility than an impious life.

Exercise 1.10 ⋆⋆

Has Pascal identified the acts, states, and outcomes correctly? If not, what did he get wrong?

1.5 The nature of belief and desire

A major obstacle to the systematic study of belief and desire is the apparent familiarity of the objects. We all know what beliefs and desires are; we have been thinking and talking about them from an early age and continue to do so almost every day. We may sometimes ask how a peculiar belief or unusual desire came about, but the nature and existence of the states themselves seems unproblematic. It takes effort to appreciate what philosophers call **the problem of intentionality**: the problem of explaining *what makes it the case* that an agent has certain beliefs and desires.

For example, some people believe that there is life on other planets, others don't. What accounts for this difference? Presumably the difference between the two kinds of people can be traced to some difference in their brains, but what is that difference, and how does a certain wiring and chemical activity between nerve cells constitute a belief in alien life? More vividly, what would you have to do in order to create an artificial agent with a belief in alien life? (Notice that producing the sounds 'there is life on other planets' is neither necessary nor sufficient.)

If we allow for degrees of belief and desire (as we should), the problem of intentionality takes on a slightly different form: what makes it the case that an agent has a belief or desire *with a given degree*? For example, what makes it the case that my credence in the existence of alien life is greater than 0.5? What makes it the case that I give greater utility to sleeping in bed than to sleeping on the floor?

These may sound like obscure philosophical questions, but they turn out to be crucial for a proper assessment of the models we will study.

I already mentioned that economists often identify utility with personal wealth or welfare. On that interpretation, the MEU Principle says that rational agents are guided solely by the expected amount of personal wealth or welfare associated with various outcomes. Yet most of us would readily sacrifice some amount of wealth or welfare in order to save a child drowning in a pond. Are we thereby violating the MEU Principle?

In general, we can't assess the MEU Principle unless we have some idea of how utility and credence (and thereby expected utility) are to be understood. There is a lot of cross-talk in the literature because authors tacitly interpret these terms in slightly different ways.

So to put flesh on the MEU Principle, we will have to say more about what we mean by 'credence' and 'utility'. I have informally introduced credence as degree of belief, and utility as degree of desire, but we should not assume that the mental vocabulary we use in everyday life precisely carves our objects of study at their joints.

For example, the word 'desire' sometimes suggests an unreflective propensity or aversion. In that sense, rational agents often act against their desires, as when I refrain from eating a fourth slice of cake, knowing that I will feel sick afterwards. By contrast, an agent's utilities comprise everything that matters to the agent – everything that motivates them, from bodily cravings to moral principles. It does not matter whether we would ordinarily call these things 'desires'.

The situation we here face is ubiquitous in science. Scientific theories often involve expressions that are given a special, technical sense. Newton's laws of motion, for example, speak of 'mass' and 'force'. But Newton did not use these words in their ordinary sense; nor did he explicitly give them a new meaning: he nowhere defines 'mass' and 'force'. Instead, he tells us what these things *do*: objects accelerate at a rate equal to the ratio between the force acting upon them and their mass, and so on. These laws implicitly define the Newtonian concept of mass and force.

We will adopt a similar perspective towards credence and utility. That is, we won't pretend that we have a perfect grip on these quantities from the outset. Instead, we'll start with a vague and intuitive conception of credence and utility and then successively refine this conception as we develop our model.

One last point. I emphasize that we are studying a **model** of belief, desire, and rational choice. Outside fundamental physics, models always involve simplifications and idealisations. In that sense, "all models are wrong", as the statistician George Box once put it. The aim of scientific models (outside fundamental physics) is not to provide a complete and fully accurate description of certain events in the world – be it the diffusion of gases, the evolution of species, or the relationship between interest rates and inflation – but to isolate simple and robust patterns in these events. It is not an objection to a model that it leaves out details or fails to explain various edge cases.

The model we will study is an extreme case insofar as it abstracts away from

most of the contingencies that make human behaviour interesting. Our topic is not specifically human behaviour and human cognition, but what unifies all types of rational behaviour and cognition.

1.6 Further reading

The use of decision matrices, dominance reasoning, and the MEU Principle are best studied through examples. A good starting point is the Stanford Encyclopedia entry on Pascal's Wager, which carefully dissects exercise 1.9:

- Alan Hájek: Pascal's Wager (2017)

Some general rules for how to identify the right acts, states, and outcomes can be found in

- James Joyce: "Decision Problems", chapter 2 of *The Foundations of Causal Decision Theory* (1999)

We will have a lot more to say about credence, utility, and the MEU Principle in later chapters. You may find it useful to read up on modelling in general and on the "functionalist" conception of beliefs and desires. Some recommendations:

- Ansgar Beckermann: "Is there a problem about intentionality?" (1996)
- Alisa Bokulich: "How scientific models can explain" (2011)
- Mark Colyvan: "Idealisations in normative models" (2013)

Essay Question 1.1

Rational agents proportion their beliefs to their evidence. Evidence is what an agent learns through perception. So could we just as well explain rational choice on the basis of an agent's *perceptions* and desires rather than her *beliefs* and desires?

2 Belief as Probability

2.1 Subjective and objective probability

Beliefs vary in strength. I believe that the 37 bus goes to Waverley station, and that there are busses from Waverley to the airport, but the second belief is stronger than the first. With some idealization, we can imagine that for any propositions A and B, a rational agent is either more confident in A than in B, more confident in B than in A, or equally confident of both. The agent's belief state then effectively sorts the propositions from 'least confident' to 'most confident', and we can represent a proposition's place in the ordering by a number between 0 ('least confident') and 1 ('most confident'). This number is the agent's degree of belief, or **credence**, in the proposition. For example, my credence in the proposition that the 37 bus goes to Waverley might be around 0.8, while my credence in the proposition that there are busses from Waverley to the airport is around 0.95.

The core assumption that unifies "Bayesian" approaches to epistemology, statistics, decision theory, and other areas, is that rational degrees of belief obey the formal rules of the probability calculus. For that reason, degrees of belief are also called **subjective probabilities** or even just **probabilities**. But this terminology can gives rise to confusion because the word 'probability' has other, and more prominent, uses.

Textbooks in science and statistics often define probability as relative frequency. On that usage, the probability of some outcome is the proportion of that type of outcome in some base class of events. For example, on the textbook definition, what it means to say that the probability of getting a six when throwing a regular die is $1/6$ is that the proportion of sixes in a large class of throws is (or converges to) $1/6$.

Another use of 'probability' is related to determinism. Consider a particular die in mid-roll. Could one in principle figure out how the die will land, given full information about its present physical state, the surrounding air, the surface on which it rolls, and so on? If yes, there's a sense in which the outcome is not a

matter of probability. Quantum physics seems to suggest that the answer is no: that the laws of nature together with the present state of the world only fix a certain probability for future events. This kind of probability is sometimes called 'chance'.

Chance and relative frequency are examples of **objective probability**. Unlike degrees of belief, they are not relative to an agent; they don't vary between you and me. You and I may have different opinions about chances or relative frequencies; but that would just be an ordinary disagreement. At least one of us would be wrong. By contrast, if you are more confident that the die will land six than me, then your subjective probability for that outcome really is greater than mine.

In this course, when we talk about credences or subjective probabilities, we do not mean beliefs about objective probability. We simply mean degrees of belief. I emphasize this point because there is a tendency, especially among economists, to interpret the probabilities in expected utility theory as objective probabilities. On that view, the MEU Principle only holds for agents who know the objective probabilities. On the (Bayesian) approach we will take instead, the MEU Principle does not presuppose knowledge of objective probabilities; it only assumes that the agent in question has a definite degree of belief in the relevant states.

2.2 Probability theory

What all forms of probability, objective and subjective, have in common is a certain abstract structure, which is studied by the mathematical discipline of probability theory.

Mathematically, a **probability measure** is a certain kind of function (in the mathematical sense, i.e. a mapping) from some objects to real numbers. The objects that are mapped to numbers are usually called 'events', but in philosophy we call them **propositions**.

The main assumption probability theory makes about propositions (the objects that are assigned probabilities) is the following.

Booleanism

Whenever some proposition A has a probability, then so does its negation $\neg A$ ('not A'). Whenever two propositions A and B both have a probability, then so does their conjunction $A \wedge B$ ('A and B') and their disjunction $A \vee B$ ('A or B').

The Bayesian approach to belief therefore implies that if a rational agent has a definite degree of belief in some propositions, then she also has a definite degree of belief in any proposition that can be construed from the original propositions in terms of negation, conjunction, and disjunction. Having a degree of belief in a proposition therefore shouldn't be understood as making a conscious judgement about the proposition. If you judge that it's likely to rain and unlikely to snow, you don't thereby make a conscious judgement about, say, $rain \vee (\neg rain \wedge snow)$.

What sorts of things are propositions? Probability theory doesn't say. In line with our discussion in the previous chapter, we will informally understand propositions as possible states of the world. This is not a formal definition, since I won't define the concept of a possible state of the world. But I'll make a few remarks that should help clarify what I have in mind.

Different sentences can represent the very same state of the world. For example, I don't know what that current temperature is in Edinburgh; one possibility (one possible state of the world) is that it is 10°C. How is this possibility related to the possibility that it is 50°F? Since 10°C = 50°F, the second possibility is not an *alternative* to the first. It is the very same possibility, expressed in a different unit. The sentences 'It is 10°C' and 'It is 50°F' are different ways of picking out the same possible state of the world.

Like sentences, possible states of the world can be negated, conjoined, and disjoined. The negation of the possibility that it is 10°C is the possibility that it is *not* 10°C. If we negate that negated state, we get back the original state: the possibility that it is *not not* 10°C coincides with the possibility that it is 10°C. In general, if we understand propositions as possible states of the world, then logically equivalent propositions are not just equivalent, but identical.

Possible states of the world can be more or less specific. That the temperature is 10°C is more specific than that it is between 7°C and 12°C. It is often useful to think of unspecific states as sets of more specific states. We can think of the possibility that it is between 7°C and 12°C as a collection of several possibilities: { 7°C, 8°C, 9°C, 10°C, 11°C, 12°C }, or even more if we consider fractional values. The unspecific possibility obtains just in case one of the more specific possibilities obtains. The most specific states are also known as **possible worlds** (in philosophy, and as 'outcomes' in most other disciplines). So we'll sometimes model propositions as sets of possible worlds.

I should warn that the word 'proposition' has many uses in philosophy. In this course, all we mean by 'proposition' is 'object of credence'. And 'credence', recall, is a semi-technical term for a certain quantity in the model we are building. It is

pointless to argue over the nature of propositions before we have spelled out the model in more detail. Also, by 'possible world' I just mean 'maximally specific proposition'. The identification of propositions with sets of possible worlds is not supposed to be an informative reduction.

Exercise 2.1 ⋆⋆

First a reminder of some terminology from set theory: The **intersection** of two sets A and B is the set of objects that are in both A and B. The **union** of two sets A and B is the set of objects that are in one or both of A and B. The **complement** of a set A is the set of objects that are not in A. A set A is a **subset** of a set B if all objects in A are also in B.

Now, assume propositions are modelled as sets of possible worlds. Then the negation $\neg A$ of a proposition A is the complement of A.

(a) What is the conjunction $A \wedge B$ of two propositions, in set theory terms?
(b) What is the disjunction $A \vee B$?
(c) What does it mean if a proposition A is a subset of a proposition B?

Exercise 2.2 ⋆⋆

The objects of probability can't all be construed as possible states of the world: it follows from Booleanism that at least one object of probability is always *impossible*. Can you explain why?

Let's return to probability theory. I said a probability measure is a function from propositions to numbers that satisfies certain conditions. These conditions are called **probability axioms** or **Kolmogorov axioms**, because their canonical statement was presented in 1933 by the Russian mathematician Andrej Kolmogorov.

The Kolmogorov Axioms

(i) For any proposition A, $0 \leq \mathrm{Cr}(A) \leq 1$.
(ii) If A is logically necessary, then $\mathrm{Cr}(A) = 1$.
(iii) If A and B are logically incompatible, then $\mathrm{Cr}(A \vee B) = \mathrm{Cr}(A) + \mathrm{Cr}(B)$.

Here I've used 'Cr' as the symbol for the probability measure, as we'll be mostly interested in subjective probability or credence. '$\mathrm{Cr}(A)$' should be read as 'the

subjective probability of A' or 'the credence in A'. Strictly speaking, we should add subscripts, '$\mathrm{Cr}_{i,t}(A)$', to make clear that subjective probability is relative to an agent i and a time t; but since we're mostly dealing with rules that hold for all agents at all times (or the relevant agent and time is clear from context), we will usually omit the subscripts.

Understood as a condition on rational credence, axiom (i) says that credences range from 0 to 1: you can't have a degree of belief greater than 1 or less than 0. Axiom (ii) says that if a proposition is logically necessary – like *it is raining or it is not raining* – then it must have subjective probability 1. Axiom (iii) says that the credence in a disjunction should equal the sum of the credence in the two disjuncts, provided these are logically incompatible, meaning they can't be true at the same time. For example, since it can't be both 8°C and 12°C, your credence in the disjunctive proposition 8°C ∨ 12°C must be $\mathrm{Cr}(8°\mathrm{C}) + \mathrm{Cr}(12°\mathrm{C})$.

We'll ask about the justification for these assumptions later. First, let's derive a few consequences.

2.3 Some rules of probability

Suppose your credence in the hypothesis that it is 8°C is 0.3. Then what should be your credence in the hypothesis that it is *not* 8°C? Answer: 0.7. In general, the probability of $\neg A$ is always 1 minus the probability of A:

The Negation Rule
$\mathrm{Cr}(\neg A) = 1 - \mathrm{Cr}(A)$.

This follows from the Kolmogorov axioms. Here is the proof. Let A be any proposition. Then $A \vee \neg A$ is logically necessary. So by axiom (ii),

$$\mathrm{Cr}(A \vee \neg A) = 1$$

Moreover, A and $\neg A$ are logically incompatible. So by axiom (iii),

$$\mathrm{Cr}(A \vee \neg A) = \mathrm{Cr}(A) + \mathrm{Cr}(\neg A)$$

Combining these two equations yields

$$1 = \mathrm{Cr}(A) + \mathrm{Cr}(\neg A)$$

From that, simple algebraic rearrangement give us the Negation Rule.

Next, we can prove that logically equivalent propositions always have the same probability.

The Equivalence Rule

If A and B are logically equivalent, then $\mathrm{Cr}(A) = \mathrm{Cr}(B)$.

Proof: Assume A and B are logically equivalent. Then $A \vee \neg B$ is logically necessary; so by axiom (ii),

$$\mathrm{Cr}(A \vee \neg B) = 1.$$

Moreover, A and $\neg B$ are logically incompatible, so by axiom (iii),

$$\mathrm{Cr}(A \vee \neg B) = \mathrm{Cr}(A) + \mathrm{Cr}(\neg B).$$

By the Negation Rule,

$$\mathrm{Cr}(\neg B) = 1 - \mathrm{Cr}(B).$$

Putting all this together, we have

$$1 = \mathrm{Cr}(A) + 1 - \mathrm{Cr}(B).$$

Subtracting $1 - \mathrm{Cr}(B)$ from both sides yields $\mathrm{Cr}(A) = \mathrm{Cr}(B)$.

Above I mentioned that if we understand propositions as possible states of the world, then logically equivalent propositions are identical: $\neg\neg A$, for example, is the very same proposition as A. The Equivalence Rule shows that even if we had used a different conception of propositions that allows distinguishing between logically equivalent propositions, these differences would never matter to an agent's subjective probabilities. If an agent's credences satisfy the Kolmogorov axioms, then she must give the same credence to logically equivalent propositions.

Exercise 2.3 ⋆⋆⋆

Prove from Kolmogorov's axioms that $\mathrm{Cr}(A) = \mathrm{Cr}(A \wedge B) + \mathrm{Cr}(A \wedge \neg B)$. (Like the proofs above, each step of your proof should either be an instance of an axiom, or an application of the rules we have already established, or it should follow from earlier steps by simple logic and algebra.)

Next, let's show that axiom (iii) generalizes to three disjuncts:

Additivity for three propositions

If A, B, and C are all incompatible with one another, then $\mathrm{Cr}(A \vee B \vee C) = \mathrm{Cr}(A) + \mathrm{Cr}(B) + \mathrm{Cr}(C)$.

Proof sketch: $A \vee B \vee C$ is equivalent (or identical) to $(A \vee B) \vee C$. If A, B, and C are mutually incompatible, then $A \vee B$ is incompatible with C. So by axiom (iii), $\mathrm{Cr}((A \vee B) \vee C) = \mathrm{Cr}(A \vee B) + \mathrm{Cr}(C)$. Again by axiom (iii), $\mathrm{Cr}(A \vee B) = \mathrm{Cr}(A) + \mathrm{Cr}(B)$. Putting these together, we have $\mathrm{Cr}((A \vee B) \vee C) = \mathrm{Cr}(A) + \mathrm{Cr}(B) + \mathrm{Cr}(C)$.

The argument generalizes to any finite number of propositions $A, B, C, D, \ldots$: the probability of a disjunction of n mutually incompatible propositions is the sum of the probability of the n propositions. This has the following consequence, which is worth remembering:

Probabilities from worlds

If the number of possible worlds is finite, then the probability of any proposition is the sum of the probability of the worlds at which the proposition is true.

Suppose two dice are tossed. There are 36 possible outcomes ("possible worlds"), which we might tabulate as follows.

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Suppose you give equal credence $1/36$ to each of these outcomes. What credence should you then give to the hypothesis that both dice land on a number less than 4? Looking at the table, we can see that there are nine possible worlds at which the hypothesis is true: the top left quarter of the table. The hypothesis is equivalent to the *disjunction* of these possible worlds. Both dice land on a number less than 4 iff the outcome is (1,1) *or* (1,2) *or* (1,3) *or* (2,1) *or* (2,2) *or* (2,3) *or* (3,1) *or* (3,2) *or* (3,3). All of these outcomes are incompatible with one another. (For example, the dice can't land (1,1) and (1,2) at the same time.) The rules of probability therefore tell us that the probability of our target hypothesis is the sum of the probability of the individual worlds. Since each world has probability $1/36$, and there are nine relevant worlds, your credence that both dice land on a number less then 4 should therefore be $9 \cdot 1/36 = 9/36 = 1/4$.

Exercise 2.4 ⋆

What credence should you give to the following propositions, in the scenario with the two dice?

(a) At least one die lands 6.
(b) Exactly one die lands 6.
(c) The sum of the numbers that will come up is equal to 5.

What if there are infinitely many worlds? Then things become tricky. It would be nice if we could say that the probability of a proposition is always the sum of the probability of the worlds that make up the proposition, but if there are too many worlds, this turns out to be incompatible with the mathematical structure of the real numbers. The most one can safely assume is that the additivity principle holds if the number of worlds is *countable*, meaning that there are no more worlds than there are natural numbers 1,2,3,.... To secure this, axiom (iii) – which is known as the axiom of **Finite Additivity** – has to be replaced by an axiom of **Countable Additivity**. In this course, we will try to stay away from troubles arising from infinities, so for our purposes, the weaker axiom (iii) will be enough.

Exercise 2.5 ⋆⋆⋆

Prove from Kolmogorov's axioms that if A entails B, then $\mathrm{Cr}(A)$ can not be greater than $\mathrm{Cr}(B)$. (You may use the rules we have already derived. Hint: if A entails B, then A is equivalent to $A \wedge B$.)

2.4 Conditional credence

To continue, we need two more concepts. The first is the idea of **conditional probability** or, more specifically, **conditional credence**. Intuitively, an agent's conditional credence reflects her degree of belief in a given proposition on the supposition that some other proposition is true. For example, I am fairly confident that it won't snow tomorrow, and that the temperature will be above 4°C. But on the supposition that it will snow, I am not at all confident that the temperature will be above 4°C. So my *unconditional credence* in temperatures above 4°C is high, while my *conditional credence* in the same proposition, on the supposition that it will snow, is low.

Conditional credence relates two propositions: the proposition that is supposed,

and the proposition that gets evaluated on the basis of that supposition.

To complicate things, there are actually two kinds of supposition, and two kinds of conditional credence. The two kinds of supposition correspond to a grammatical distinction between "indicative" and "subjunctive" conditionals. Compare the following pair of statements.

(1) If Shakespeare didn't write *Hamlet*, then someone else did.

(2) If Shakespeare hadn't written *Hamlet*, then someone else would have.

The first of these (an indicative conditional) is highly plausible: we know that someone wrote *Hamlet*; if it wasn't Shakespeare then it must have been someone else. By contrast, the second statement (a subjunctive conditional) is plausibly false: if Shakespeare hadn't written *Hamlet*, it is unlikely that somebody else would have stepped in to write the very same play.

The two conditionals (1) and (2) relate the same two propositions – the same possible states of the world. To evaluate either statement, we suppose that the world is one in which Shakespeare didn't write *Hamlet*. The difference lies in what we hold fixed when we make that supposition. To evaluate (1), we hold fixed our knowledge that *Hamlet* (the play) exists. Not so in (2). To evaluate (2), we bracket everything we know that we take to be a causal consequence of Shakespeare's writing of *Hamlet*.

We will return to the second, subjunctive kind of supposition later. For now, let's focus on the first, indicative kind of supposition. We will write $\mathrm{Cr}(A/B)$ for the **(indicative) conditional credence** in A on the supposition that B. Again, intuitively this is the agent's credence that A is true *if* (or *given that* or *supposing that*) B is true.

How are conditional credences related to unconditional credences? The answer is surprisingly simple, and captured by the following formula.

The Ratio Formula

$$\mathrm{Cr}(A/B) = \frac{\mathrm{Cr}(A \wedge B)}{\mathrm{Cr}(B)}, \text{ provided } \mathrm{Cr}(B) > 0.$$

That is, your credence in some proposition A on the (indicative) supposition B equals the ratio of your unconditional credence in $A \wedge B$ divided by your unconditional credence in B.

To see why this makes sense, it may help to imagine your credence as distributing a certain quantity of "plausibility mass" over the space of possible worlds.

When we ask about your credence in A conditional on B, we set aside worlds where B is false. What we want to know is how much of the mass given to B worlds falls on A worlds. In other words, we want to know what fraction of the mass given to B worlds is given to A worlds that are also B worlds.

People disagree on the status of the Ratio Formula. Some treat it as a definition. On that approach, you can ignore everything I said about what it means to suppose a proposition and simply read '$\text{Cr}(B/A)$' as shorthand for '$\text{Cr}(A \wedge B)/\text{Cr}(A)$'. Others regard conditional beliefs as distinct and genuine mental states and see the Ratio Formula as a fourth axiom of probability. We don't have to adjudicate between these views. What matters is that the Ratio Formula is true, and on this point both sides agree.

The second concept I want to introduce is that of probabilistic independence. We say that propositions A and B are **(probabilistically) independent** (for the relevant agent at the relevant time) if $\text{Cr}(A/B) = \text{Cr}(A)$. Intuitively, if A and B are independent, then it makes no difference to your credence in A whether or not you suppose B, so your unconditional credence in A is equal to your credence in A conditional on B.

Note that unlike causal independence, probabilistic independence is a feature of beliefs. It can easily happen that two propositions are independent for one agent but not for another. That said, there are interesting connections between probabilistic (in)dependence and causal (in)dependence. For example, if an agent knows that two events are causally independent, then the events are normally also independent in the agent's degrees of belief. You may want to ponder why that is the case.

Exercise 2.6 ⋆

Assume $\text{Cr}(\mathit{Snow}) = 0.3$, $\text{Cr}(\mathit{Wind}) = 0.6$, and $\text{Cr}(\mathit{Snow} \wedge \mathit{Wind}) = 0.2$. What is $\text{Cr}(\mathit{Snow}/\mathit{Wind})$? What is $\text{Cr}(\mathit{Wind}/\mathit{Snow})$?

Exercise 2.7 ⋆⋆

Using the Ratio Formula, show that if A is (probabilistically) independent of B, then B is independent of A.

Exercise 2.8 ⋆⋆

A fair die will be tossed, and you give equal credence to all six outcomes. Let A

be the proposition that the die lands either 1 or 6. Let B be the proposition that the die lands an odd number (1,3, or 5). Let C be the proposition that the die lands 1, 2 or 3. Which of the following are true, in your belief state?

(a) A is independent of B.
(b) A is independent of C.
(c) A is independent of $B \wedge C$.
(d) B is independent of C.

2.5 Some more rules of probability

If you've studied propositional logic, you'll know how to compute the truth-value of arbitrarily complex sentences from the truth-value of their atomic parts. For example, if p and q are true and r is false, then you can figure out whether $\neg(p \wedge (q \vee \neg(r \vee p)))$ is true. Now suppose instead of the truth-value of p, q, and r, I give you their probability. Could you then compute the probability of $\neg(p \wedge (q \vee \neg(r \vee p)))$? The answer is no. In general, while the probability of $\neg A$ is determined by the probability of A (as we know from the Negation Rule), neither the probability of $A \vee B$ nor the probability of $A \wedge B$ is determined by the individual probabilities of A and B.

Let's have a look at conjunctive propositions, $A \wedge B$. By rearranging the Ratio Formula, we get the following:

The Conjunction Rule

$\mathrm{Cr}(A \wedge B) = \mathrm{Cr}(A) \cdot \mathrm{Cr}(B/A)$.

So the probability of a conjunction is the probability of the first conjunct times the probability of the second *conditional on the first*. If you only know the unconditional probabilities of the conjuncts, you can't figure out the probability of the conjunction.

But there's a special case. If A and B are independent, then $\mathrm{Cr}(B/A) = \mathrm{Cr}(B)$; so in that case the probability of the conjunction is the product of the probability of the conjuncts:

The Conjunction Rule for independent propositions

If A and B are independent, then $\mathrm{Cr}(A \wedge B) = \mathrm{Cr}(A) \cdot \mathrm{Cr}(B)$.

Why do we multiply (rather than, say, add) the probabilities in the Conjunction Rules? Suppose we flip two coins. What is the probability that they both land heads? You'd expect the first coin to land heads about half the time; and in half *of those* cases you'd expect the second to also land heads. So the result is a half of a half; that is, $1/2 \cdot 1/2 = 1/4$.

What about disjunctions, $A \vee B$? We know that if A and B are logically incompatible, then $\text{Cr}(A \vee B) = \text{Cr}(A) + \text{Cr}(B)$. But what if A and B are not incompatible? In that case, we have to subtract the probability of the conjunction:

The Disjunction Rule

$Cr(A \vee B) = Cr(A) + Cr(B) - Cr(A \wedge B)$.

Again, you can't compute the probability of the disjunction just from the probability of the disjuncts.

If you want to know where the Disjunction Rule comes from, note that in general, the proposition $A \vee B$ comprises three kinds of worlds: (1) worlds where A is true and B is false, (2) worlds where B is true and A is false, and (3) worlds where A and B are both true. These three sets are disjoint (mutually exclusive). So by Additivity, the probability of $A \vee B$ is equal to the probability of $A \wedge \neg B$ plus the probability of $B \wedge \neg A$ plus the probability of $A \wedge B$. Taken together, the worlds in (1) and (3) comprise precisely the A-worlds, and the worlds in (2) and (3) comprise the B-worlds. So if we add together $\text{Cr}(A)$ and $\text{Cr}(B)$, we have effectively double-counted the $A \wedge B$ worlds. So we need to subtract $\text{Cr}(A \wedge B)$.

Exercise 2.9 ⋆

Show that two propositions A and B are independent if and only if $Cr(A \wedge B) = Cr(A) \cdot Cr(B)$. (Some authors use this as the definition of independence.)

Exercise 2.10 ⋆⋆

Prove from the Ratio Formula that $\text{Cr}(A \wedge B \wedge C) = \text{Cr}(A/B \wedge C) \cdot \text{Cr}(B/C) \cdot \text{Cr}(C)$.

Exercise 2.11 ⋆

In 1999, a British woman was convicted of the murder of her two sons, who she claimed died from Sudden Infant Death Syndrome (SIDS). The eminent paediatrician Sir Roy Meadow explained to the jury that 1 in 8500 infants die

from SIDS and hence the chance of SIDS affecting both sons was $1/8500 \cdot 1/8500 = 1$ in 73 million. What is wrong with Sir Meadow's reasoning?

To conclude, I'll mention two more rules that play a special role in Bayesian approaches to belief. The first goes back to a suggestion by Thomas Bayes published in 1763.

Bayes' Theorem

$$\mathrm{Cr}(A/B) = \frac{\mathrm{Cr}(B/A) \cdot \mathrm{Cr}(A)}{\mathrm{Cr}(B)}$$

Proof: By the Ratio Formula, $\mathrm{Cr}(A/B) = \mathrm{Cr}(A \wedge B)/\mathrm{Cr}(B)$. By the Conjunction Rule, $\mathrm{Cr}(A \wedge B) = \mathrm{Cr}(B/A) \cdot \mathrm{Cr}(A)$. So we can substitute $\mathrm{Cr}(A \wedge B)$ in the Ratio Formula by $\mathrm{Cr}(B/A) \cdot \mathrm{Cr}(A)$, which yields Bayes' Theorem.

Bayes' Theorem relates the conditional credence in A given B to the inverse conditional credence in B given A. Why that might be useful is best illustrated by an example.

Suppose you are unsure whether the die I am about to roll is a regular die or a trick die that has a six printed on all sides. You currently give equal credence to both possibilities. How confident should you be that the die is a trick die *given that it will land six on the next roll*? That is, what is $\mathrm{Cr}(\mathit{Trick}/\mathit{Six})$? The answer isn't obvious. Bayes' Theorem helps. By Bayes' Theorem,

$$\mathrm{Cr}(\mathit{Trick}/\mathit{Six}) = \frac{\mathrm{Cr}(\mathit{Six}/\mathit{Trick}) \cdot \mathrm{Cr}(\mathit{Trick})}{\mathrm{Cr}(\mathit{Six})}.$$

The numerator on the right is easy. $\mathrm{Cr}(\mathit{Six}/\mathit{Trick})$ is 1: if the die has a six on all its sides then it is certain that it will land six. We also know that $\mathrm{Cr}(\mathit{Trick})$ is $1/2$. But what is $\mathrm{Cr}(\mathit{Six})$, your unconditional credence that the die will land six? Here we need one last rule:

The Law of Total Probability

$$\mathrm{Cr}(A) = \mathrm{Cr}(A/B) \cdot \mathrm{Cr}(B) + \mathrm{Cr}(A/\neg B) \cdot \mathrm{Cr}(\neg B).$$

This "law" follows immediately from exercise 2.3 and the Conjunction Rule.

If we apply the Law of Total Probability to $\mathrm{Cr}(\mathit{Six})$ in the above application of Bayes' Theorem, we get

$$\mathrm{Cr}(\mathit{Trick}/\mathit{Six}) = \frac{\mathrm{Cr}(\mathit{Six}/\mathit{Trick}) \cdot \mathrm{Cr}(\mathit{Trick})}{\mathrm{Cr}(\mathit{Six}/\mathit{Trick}) \cdot \mathrm{Cr}(\mathit{Trick}) + \mathrm{Cr}(\mathit{Six}/\neg\mathit{Trick}) \cdot \mathrm{Cr}(\neg\mathit{Trick})}.$$

It looks scary, but all the terms on the right are easy to figure out. We already know that $\text{Cr}(\mathit{Six}/\mathit{Trick}) = 1$ and that $\text{Cr}(\mathit{Trick}) = 1/2$. Moreover, plausibly $\text{Cr}(\mathit{Six}/\neg\mathit{Trick}) = 1/6$ and $\text{Cr}(\neg\mathit{Trick}) = 1/2$. Plugging in all these values, we get $\text{Cr}(\mathit{Trick}/\mathit{Six}) = 6/7$. So your credence in the trick die hypothesis conditional on seeing a six should be $6/7$.

Exercise 2.12 ⋆⋆⋆

A stranger tells you that she has two children. You ask if at least one of them is a boy. The stranger says yes. How confident should you be that the other child is also a boy? (Assume both sexes are equally common and independent among siblings.)

2.6 Further reading

There are many good introductions to probability theory – for example, chapters 3–7 in

- Ian Hacking: *An Introduction to Probability and Inductive Logic* (2001).

You may also find the rest Hacking's book useful to supplement later parts of this course.

A good introduction to the problems that arise if one tries to extend Additivity to infinite cases is

- Brian Skyrms: "Zeno's paradox of measure" (1983)

If you want to tackle the essay question below, have a look at

- Robert Stalnaker: "The problem of logical omniscience I" (1991)

Essay Question 2.1

By Kolmogorov's axiom (ii), logically necessary propositions have probability 1. If an agent's degrees of belief satisfy the probability axioms, it seems to follow that the agent must be certain of every logical truth. Does this show that Bayesian models are inapplicable to real agents who are not logically omniscient? Carefully explain your answer.

3 Probabilism

3.1 Justifying the probability axioms

The hypothesis that rational degrees of belief satisfy the mathematical conditions on a probability measure is known as **probabilism**. In this chapter, we will look at some arguments for probabilism. We do so not because the hypothesis is especially controversial (by philosophy standards, it is not), but because it is instructive to reflect on how one could argue for an assumption like this, and also because the task will bring us back to a more fundamental question: what it means to say that an agent has such-and-such degrees of belief in the first place.

We will assume without argument that if a rational agent has degrees of belief in some propositions A and B, then she also has degrees of belief in their conjunction, disjunction, and negation. Probabilism then reduces to the hypothesis that rational degrees of belief satisfy the probability axioms – specifically, Kolmogorov's axioms (i)–(iii):

(i) For any proposition A, $0 \leq \mathrm{Cr}(A) \leq 1$.

(ii) If A is logically necessary, then $\mathrm{Cr}(A) = 1$.

(iii) If A and B are logically incompatible, then $\mathrm{Cr}(A \vee B) = \mathrm{Cr}(A) + \mathrm{Cr}(B)$.

Consider axiom (i). Why should rational degrees of belief always fall in the range between 0 and 1? Why would it be irrational to believe some proposition to degree 7? The question is hard to answer unless we have some idea of what it would mean to believe a proposition to degree 7.

It is tempting to think that axiom (i) does not express a substantive norm of rationality, but a convention of representation. We have decided to represent strength of belief by numbers between 0 and 1, where 1 means absolute certainty. We could just as well have decided to use numbers between 0 and 100, or between -100 and +100. Having set up the convention to put the upper limit at 1, it doesn't make any sense to assume that an agent believes something to degree 7.

Axioms (ii) and (iii) look more substantive. It seems that we can at least imagine an agent who assigns degree of belief less than 1 to a logically necessary proposition or whose credence in a disjunction of incompatible propositions is not the sum of her credence in the individual disjuncts. Still, we need to clarify what exactly it is that we're imagining if we want to discuss whether the imagined states are rational or irrational.

For example, suppose we understand strength of belief as a certain introspectible quantity: a basic feeling of conviction people have when entertaining propositions. Axiom (ii) would then say that when agents entertain logically necessary propositions, they ought to experience this sensation with maximal intensity. It is hard to see why this should be norm of rationality. It is also hard to see why the sensation should guide an agent's choices in line with the MEU Principle, or why it should be sensitive to the agent's evidence.

So if we understand degrees of belief as measuring the intensity of a certain feeling, then the norms of Bayesian decision theory and Bayesian epistemology look implausible and inexplicable. The same is true if we understand degrees of belief as measuring some other basic psychological quantity: why should that quantity satisfy the probability axioms, guide behaviour, respond to evidence, etc.?

A more promising line of thought assumes that strength of belief is defined (perhaps in part) by the MEU Principle. On that approach, what we mean when we say that an agent has such-and-such degrees of belief is that she is (or ought to be) disposed to make certain choices. We can then assess the rationality of the agent's beliefs by looking at the corresponding choice dispositions.

Unfortunately, beliefs alone do not settle rational choices: the agent's desires or goals also play a role. The argument we are now going to look at therefore fixes an agent's goals by assuming that utility equals monetary payoff. Afterwards we will consider how that assumption could be relaxed.

3.2 The betting interpretation

It is instructive to compare degrees of belief with other numerical quantities in science. Take mass. What do we mean when we say that an object – a chunk of iron perhaps – has a mass of 2 kg? There are no little numbers written in chunks of iron, just as there are no little numbers written in the head. As with degrees of belief, there is an element of conventionality in the way we represent

masses by numbers: instead of representing the chunk's mass by the number 2, we could just as well have used a different scale on which the mass would be 2000 or 4.40925. (Appending 'kg' to the number, as opposed to 'g' or 'lb', hints at the conventional scale.)

I am not suggesting that mass itself is conventional. Whether a chunk of iron has a mass of 2 kg is, I believe, a completely objective, mind-independent matter. If there were no humans, the chunk would still have that mass. What's conventional is only the representation of masses (which are not intrinsically numerical) by numbers.

The reason why we can measure mass in numbers – and the reason why we know anything at all about mass – is that things tend to behave differently depending on their mass. The greater an object's mass, the harder the object is to lift up or accelerate. Numerical measures of mass reflect these dispositions, and can be standardized by reference to particular manifestations. For example, if we put two objects on opposite ends of a balance, the one with greater mass will go down. So we could choose a random chunk of iron, call it the "standard kilogram", and stipulate that something has a mass of n kg just in case it balances against n copies of the standard kilogram (or against n objects each of which balances against the standard kilogram).

Can we take a similar approach to degrees of belief? The idea would be to find a characteristic way in which degrees of belief manifest themselves in behaviour and use that to define a numerical scale for degrees of belief.

So how do you measure someone's degrees of belief? The classical answer is: by offering them a bet.

Consider a bet that pays £1 if it will rain at noon tomorrow, and nothing if it won't rain. How much would you be willing to pay for this bet?

We can calculate the expected payoff – that is, the average of the possible payoffs, weighted by their subjective probability. Suppose your degree of belief in rain tomorrow is x, and your degree of belief in not-rain is $1-x$. Then the bet would give you £1 with probability x and £0 with probability $1-x$. So the expected payoff is $x \cdot £1 + (1-x) \cdot £0 = £x$. This suggests that the bet is worth $£x$. That is, $£x$ is the most you should pay for the bet.

Exercise 3.1 ⋆

Suppose your degree of belief in rain is 0.8 (and your degree of belief in not-rain 0.2). For a price of £0.70 you can buy a bet that pays £1 if it rains and £0 if it

doesn't rain. Draw a decision matrix for your decision problem and compute the expected utility of the acts, assuming your subjective utilities equal the net amount of money you have gained in the end.

If we're looking for a way to measure your degrees of belief, we can turn this line of reasoning around: if £x is the most you're willing to pay for the bet, then x is your degree of belief in the proposition that it will rain. This leads to the following suggestion.

The betting interpretation

An agent believes a proposition A to degree x just in case she would pay up to £x for a bet that pays £1 if A is true and £0 otherwise.

The betting interpretation is meant to have the same status as the above (hypothetical) stipulation that an object has a mass of n kg just in case it balances against n copies of the standard kilogram. On the betting interpretation, offering people bets is like putting objects on a balance scale. For some prices, the test person will prefer to buy the bet, for others she will be prefer to sell the bet; in between there is a point at which the price of the bet is in balance with the expected payoff, so the test person will be indifferent between buying, selling, and doing neither. The price at the point of balance reveals the subject's degree of belief. The stake of £1 is a unit of measurement, much like the standard kilogram in the measurement of mass.

The betting interpretation gives us a clear grip on what it means to believe a proposition to a particular degree. It also points towards an argument for probabilism. For we can show that if an agent's degrees of belief do not satisfy the probability axioms (for short, if her beliefs are not **probabilistic**), then the agent is disposed to enter bets that amount to a guaranteed loss.

3.3 The Dutch Book theorem

In what follows, we are going to assume that if an agent is not willing to buy a bet for £x, then she would be be willing to sell the bet for that price. Selling a bet means offering it to somebody else. The idea is that if you judge a bet to worth less than £x, then you should be happy to offer someone the bet for a price of £x.

Exercise 3.2 ⋆

Suppose your degree of belief in rain is 0.8 (and in not-rain 0.2) and someone offers you £0.90 for a bet that pays £1 if it rains and £0 if it doesn't rain. Should you sell (i.e. offer) them the bet? Draw a decision matrix for your decision problem and compute the expected utility of the acts, assuming your subjective utilities equal the net amount of money you get in the end.

In betting jargon, a combination of bets (bought or sold) is called a 'book'. A combination of bets that amounts to a guaranteed loss is a **Dutch book**. We will now prove that if an agent's degrees of belief violate one or more of the Kolmogorov axioms, and she values bets in accordance with their expected payoff, then she is prepared to accept a Dutch Book.

We begin with axiom (i). Suppose your credence in some proposition A is greater than 1. For concreteness, let's say $\mathrm{Cr}(A) = 2$. By the betting interpretation, this means you'd be willing to pay up to £2 for a deal that pays you back either £0 or £1, depending on whether A is true. You're guaranteed to lose at least £1. More generally, if your degree of belief in A is greater than 1, then you are guaranteed to lose at least the difference between your degree of belief and 1.

Similarly, suppose your credence in A is below 0. Let's say it's -1. By the betting interpretation, this means you would pay no more than £-1 for a bet on A and you'd be willing to sell the bet for any price above £-1. What does it mean to sell a bet for £-1? It means to pay someone £1 to take the bet. So you would be willing to pay up to £1 for me to take the bet from you, with no chance of getting any money back. You're guaranteed to lose at least £1. Again, the argument generalizes to any degree of belief below 0.

I leave the case of axiom (ii) as an exercise.

Exercise 3.3 ⋆⋆

Suppose your degrees of belief violate axiom (ii). Assuming the betting interpretation, describe a bet you are willing to sell for a price less than what you could possibly get back.

Now for axiom (iii). Suppose your credence in the disjunction of two logically incompatible propositions A and B is not the sum of your credence in the individual propositions. For concreteness, let's assume $\mathrm{Cr}(A) = 0.4$, $\mathrm{Cr}(B) = 0.2$, and $\mathrm{Cr}(A \vee B) = 0.5$. By the betting interpretation, you'll then be willing to sell a bet on $A \vee B$ for at least £0.50, and you'll be willing to buy a bet on A for up to

£0.40 and a bet on B for up to £0.20. If you buy these two bets you have in effect bought a bet on $A \vee B$, for you will get £1 if either A or B is true, and £0 otherwise. So you are (in effect) willing to sell this bet for £0.50 and buy it back for £0.60. No matter how the bets turn out, you will lose £0.10 (as you can check).

The reasoning generalizes to any other case where $\mathrm{Cr}(A \vee B) < \mathrm{Cr}(A) + \mathrm{Cr}(B)$. For cases where $\mathrm{Cr}(A \vee B) > \mathrm{Cr}(A) + \mathrm{Cr}(B)$, simply swap all occurrences of 'buy' and 'sell' in the previous paragraph.

We have thereby proved the *Dutch Book Theorem*.

Dutch Book Theorem

If an agent values bets by their expected monetary payoff and her degrees of belief don't conform to the Kolmogorov axioms, then she is prepared to accept combinations of bets that amount to a guaranteed loss.

Note that the Dutch Book Theorem is a conditional: *if* an agent has non-probabilistic beliefs and values bets by their expected monetary payoff, *then* she is vulnerable to Dutch Books. We have not shown that agents with probabilistic beliefs are immune to Dutch Books. But that can also be shown (with some further restrictions on the relevant agents); the result is known as the **Converse Dutch Book Theorem**. I won't go through the proof.

In chapter 2, I mentioned that some authors treat the Ratio Formula for conditional probability as a definition while others treat it fourth axiom of probability. On the second perspective, we might want to show that violations of that fourth axiom also make an agent vulnerable to Dutch Books.

To this end, we would first have to extend the betting interpretation, in order to clarify how conditional credences manifest themselves in betting behaviour. The standard approach is to introduce the idea of a conditional bet. A *unit bet on A conditional on B* is a bet that only comes into effect if B is true. In that case it pays £1 if A is true and £0 if A is false. If B is not true, whoever bought the bet gets a refund for the price they paid. Now we can extend the betting interpretation to say that your conditional credence in A given B is the maximal price at which you would be willing to buy the corresponding conditional bet. And then it is not hard to show that a Dutch Book can be made against you unless your conditional credences satisfy the Ratio Formula.

Exercise 3.4 ⋆⋆⋆

Suppose I believe that it is raining to degree 0.6 and that it is not raining also to degree 0.6. Describe a Dutch Book you could make against me, assuming the betting interpretation.

3.4 Problems with the betting interpretation

The Dutch Book Theorem is a mathematical result. It does not show that rational degrees of belief satisfy the probability axioms. To reach that conclusion, and thereby an argument for probabilism, we need to add some philosophical premises about rational belief.

On a flat-footed interpretation, we might take the theorem as a warning that if our degrees of belief do not satisfy the probability axioms, then a cunning Dutchman might come along and trick us out of money. But does this really show that non-probabilistic beliefs are irrational? Two problems immediately stand out.

First, why should the mere possibility of financial loss be a sign of irrational beliefs? True, there might be a Dutchman going around exploiting people with non-probabilistic beliefs. But there might also be someone (a Frenchman, say) going around richly rewarding people with non-probabilistic beliefs. We don't think the latter possibility shows that people ought to have non-probabilistic beliefs. Even if there is such a Frenchman, we can at most conclude that it would be *practically useful* to have non-probabilistic beliefs. Arguably those beliefs would still not be *epistemically rational*. (Compare: if someone offers you a million pounds if you believe that the moon is made of cheese, then the belief would be practically useful, but it would not be epistemically justified; it would not reflect your evidence.) Why should we think differently about the hypothetical Dutchman?

Second, the threat of financial exploitation only awaits non-probabilistic agents who value bets by their expected monetary payoff and thus willing buy or sell any bet if the expected monetary payoff of the transaction is positive. This is entailed by the betting interpretation, but on reflection it is untenable.

Consider the following gamble.

Example 3.1 (The St. Petersburg Paradox)

I will toss a fair coin until it lands tails. If the coin lands tails on the first toss, you get £2. If it lands heads on the first toss and tails on the second, you get

£4. If the coin lands heads on the first two tosses and tails on the third, you get £8. And so on: if the coin first lands tails on the nth toss, you get £2^n.

How much would you pay for this gamble?

We can compute the expected payoff. With probability $1/2$ you'll get £2; with probability $1/4$ you get £4; with probability $1/8$ you get £8; and so on. The expected payoff is therefore

$$1/2 \cdot £2 + 1/4 \cdot £4 + 1/8 \cdot £8 + \ldots = £1 + £1 + £1 + \ldots.$$

The sum of this series is infinite. That is, if you value bets by their expected monetary payoff, you should sacrifice everything you have for an opportunity to play the gamble. In reality, few people would do that, seeing as the payoff is almost certain to be quite low.

Exercise 3.5 ⋆

What is the probability that you will get £16 or less when playing the St. Petersburg gamble?

When the St. Petersburg Paradox was first described by the Swiss mathematician Nicolas Bernoulli (in 1713), it motivated his cousin Daniel Bernoulli to introduce the theoretical concept of utility as distinct from monetary payoff. As (Daniel) Bernoulli realised, "a gain of one thousand ducats is more significant to the pauper than to a rich man though both gain the same amount". In other words, most people don't regard having two million pounds as twice as good as having one million pounds: the first million would make a much greater difference to our lives than the second.

In economics terminology, what Bernoulli realised is that money has **declining marginal utility**. The 'marginal utility' of a good for an agent is how much she desires an extra unit of the good. To say that the marginal utility of money is declining therefore means that the more money you have, the less you value an additional pound.

Concretely, Daniel Bernoulli suggested that n units of money provide not n but $\log(n)$ units of utility, so that doubling your wealth from £1000 to £2000 would provide the same boost in utility than doubling your wealth from £1 million to £2 million (even though the second change is much larger in absolute terms). On Bernoulli's model, the expected utility of the St. Petersburg gamble for a person with a wealth of £1000 is not infinite, but £10.95: that is the most she ought to be willing to pay.

There is clearly nothing irrational about an agent who assigns declining marginal utility to money. But then we can't assume that rational agents value bets by their expected monetary payoff.

Exercise 3.6 ⋆⋆

Suppose owning £n gives you a utility of $\log(n)$. You currently have £1. For a price of £0.40 you are offered a bet that pays £1 if it will rain tomorrow (and £0 otherwise). Your degree of belief in rain tomorrow is $1/2$. Should you accept the bet? Draw the decision matrix and compute the expected utilities. [You'll need to know that $\log(1) = 0$, $\log(1.6) \approx 0.47$, and $\log(0.6) \approx -0.51$. Apart from that you don't need to know what 'log' means.]

Exercise 3.7 ⋆

Bernoulli's logarithmic model is obviously a simplification. Suppose you want to take a bus home, but you only have £1.50 whereas the fare is £1.70. If you can't take the bus, you'll have to walk for 50 minutes through the rain. A stranger at the bus stop offers you a deal: if you give her your £1.50, she will toss a fair coin and pay you back £1.70 on heads or £0 on tails. Explain (briefly and informally) why it would be rational for you to accept the deal.

There's another reason why rational agents don't always value bets by their expected payoff, even if their subjective utility is adequately measured by monetary payoff. The reason is that buying or selling bets can alter the relevant beliefs.

For example, I am quite confident I will not buy any bets today. Should I therefore be prepared to pay close to £1 for a bet on the proposition that I don't buy any bets today? Clearly not. By buying the bet, I would render the proposition false. Given my current state of belief, the (imaginary) bet has an expected payoff close to £1; nonetheless, it would be irrational for me to buy it even for £0.10.

So rational agents don't always value bets by their expected payoff. The betting interpretation is untenable. An agent's betting dispositions may often give a good hint about their degrees of belief, but we can't simply read off degrees of belief from dispositions to buy and sell bets.

3.5 A Dutch Book argument

Given the issues raised in the previous section, can we learn anything about rational belief from the Dutch Book Theorems? Some philosophers have argued that we can't. I am a little more optimistic. But clearly the argument will have to be more complicated than one might initially have thought. Here is a sketch of one possible approach.

Consider an arbitrary agent with non-probabilistic beliefs. Call her α. We want to show that α's beliefs are epistemically irrational. We can't assume that α values bets by their expected monetary payoff. Perhaps α hates betting, or doesn't care about money. But these desires arguably don't affect the epistemic rationality of α's beliefs.

So let's imagine a counterpart β of α who is in the same epistemic state as α but doesn't hate betting; β, I hereby stipulate, values any bet she might be offered by the bet's expected monetary payoff. I also stipulate that β follows the MEU Principle.

My first philosophical premise is that *if α's belief state is epistemically rational, then so is β's*. The idea is that if you want to know whether an agent's beliefs are epistemically rational, you don't need to know anything about her desires or the way she makes choices. But that's the only (possible) difference between α and β.

As we saw at the end of the previous section, we can't assume that if β's credence in a proposition is x, then she will pay up to £x for a bet that pays £1 if A and £0 if not-A, since her credence in A may be affected by the transaction. But this problem only seems to arise for a small and special class of propositions. Let's call a proposition *stable* if it is probabilistically independent, in β's credence function, of the assumption that a bet on the proposition is bought or sold. Let's call a pair of propositions (A, B) *jointly stable* if both A and B are stable and β's credence in B is unaffected by assumptions about whether a bet on A has been bought or sold.

Now the probability axioms are supposed to be general consistency requirements on rational belief. Such requirements should plausibly be "topic-neutral": they should hold for beliefs or every kind, not just for beliefs about a special subject matter. So if the probability axioms are rational requirements for an agent's credences over stable propositions, then they should be requirements for an agent's entire credence function. This is my second premise: *if any restriction of an agent's credence function to stable propositions should satisfy Kolmogorov's*

(i), (ii), and (iii), then so should the agent's entire credence function.

To show that non-probabilistic beliefs are irrational, it is therefore enough to show that non-probabilistic beliefs towards (jointly) stable propositions are irrational. So we can assume without loss of generality that α's (and therefore β's) beliefs towards stable propositions are non-probabilistic.

It follows by the Dutch Book Theorem that β is prepared to knowingly buy and sell bets in such a way that she is guaranteed to lose money. My next premise states that it would be irrational for β to make these transactions. That is, *it is irrational for an agent whose sole aim is to maximize her profit in each transaction to knowingly and avoidably make transactions whose net effect is a guaranteed loss*. This premise relies on the Converse Dutch Book Theorem, which shows that it is possible to avoid making a sure loss.

So our hypothetical agent β is disposed to make irrational choices. Arguably (premise 4), *if an agent makes irrational choices, then she is epistemically irrational, or her desires are irrational, or her acts don't maximize expected utility*. In the case of β, we can rule out the third possibility.

Moreover (premise 5), *β's desires are not irrational*. Admittedly, her desires are strange. β would be prepared to give all she has for an opportunity to play the St. Petersburg gamble. But from a thoroughly subjective point of view, there is nothing incoherent or inconsistent about these desires.

So β is epistemically irrational. By the very first premise, it follows that α is epistemically irrational. And α was an arbitrary agent whose credences violate Kolmogorov's axioms. So rational credences are probabilistic.

The argument has a lot of premises, and many of them could be challenged. Can you think of a better argument?

> **Exercise 3.8** ⋆⋆⋆
>
> Why do I need the assumption of "joint stability" for pairs of propositions?

3.6 Comparative credence

We have seen that the betting interpretation is untenable. Many philosophers hold that degrees of belief cannot be defined in terms of an agent's behaviour, but should rather be treated as theoretical primitives. Even on that view, however, more must be said about the numerical representation of credence. That we represent degrees of belief by numbers between 0 and 1 is clearly a matter of

convention: whatever is represented by these numbers could just as well be represented by numbers between 0 and 100, by the rotation of a line, or in various other ways. So we need to explain the convention of assigning numbers to whatever an agent's credence function represents.

One approach towards such an explanation, which does not turn on an agent's behaviour, was outlined by the Italian mathematician and philosopher Bruno de Finetti in the 1930s. De Finetti suggested that degrees of belief could be defined in terms of the comparative attitude of being more confident in one proposition than in another. While any numerical representation of beliefs is partly conventional, this comparative attitude is plausibly objective and might be taken as primitive.

Let '$A \succ B$' express that a particular (not further specified) agent is more confident in A than in B. For example, if you are more confident that it is sunny than that it is raining, then $\mathit{Sunny} \succ \mathit{Rainy}$. Let '$A \sim B$' mean that the agent is equally confident in A and in B. From these, we can define a third relation '$\succsim$' by stipulating that $A \succsim B \iff (A \succ B) \vee (A \sim B)$.

What can we assume about the formal structure of these relations? First of all, if you're more confident in A than B, then you can't at the same time be more confident in B than A or equally confident in the two propositions. Moreover, if you're neither more confident in A than B, nor in B than A, then you're plausibly equally confident in the two. So we may assume that an agent's comparative credence relations are "complete" in the following sense:

Completeness

For any A and B, exactly one of $A \succ B$, $B \succ A$, or $A \sim B$ is the case.

Next, suppose you are at least as confident in A as in B, and at least as confident in B as in C. Then you should be at least as confident in A as in C. In other words, $\succsim$ should be "transitive":

Transitivity

If $A \succsim B$ and $B \succsim C$ then $A \succsim C$.

Exercise 3.9 ⋆⋆⋆

Show that Transitivity and Completeness together entail that (a) if $A \sim B$ then $B \sim A$, and (b) if $A \sim B$ and $B \sim C$, then $A \sim C$.

For the next assumptions, I use '$\top$' to stand for the logically necessary proposition (the set of all worlds) and '$\bot$' for the logically impossible proposition (the empty set).

Normalization
$\top \succ \bot$.

Boundedness
There is no proposition A such that $\bot \succ A$.

These are fairly plausible as demands of rationality.

The next assumption is best illustrated by an example. Suppose you are more confident that Bob is German than that he is French. Then you should also be more confident that Bob is *either German or Russian* than that he is *either French or Russian*. Conversely, if you are more confident that he is German or Russian than that he is French or Russian, then you should be more confident that he is German than that he is French. In general:

Quasi-Additivity
If A and B are both logically incompatible with C, then $A \succsim B$ iff $(A \vee C) \succsim (B \vee C)$.

De Finetti conjectured that whenever an agent's comparative credence relations satisfy the above five assumptions, then there is a unique probability measure Cr such that $A \succ B$ iff $\text{Cr}(A) > \text{Cr}(B)$ and $A \sim B$ iff $\text{Cr}(A) = \text{Cr}(B)$. The conjecture turned out to be false, because a sixth assumption is required. But the following can be shown:

Probability Representation Theorem
If an agent's comparative credence relations satisfy Completeness, Transitivity, Normalization, Boundedness, Quasi-Additivity, and the Sixth Assumption, then there is a unique probability measure Cr such that $A \succsim B$ iff $\text{Cr}(A) \geq \text{Cr}(B)$.

Before I describe the Sixth Assumption, let me explain what the Probability Representation Theorem might do for us.

I have argued that we can't take numerical credences as unanalysed primitives. There must be an answer to the question why an agent's degree of belief in rain is correctly represented by the number 0.2 rather than, say, 0.3. De Finetti's idea was to derive numerical representations of belief from comparative attitudes

towards propositions.

Imagine we order all propositions on a line, in accordance with the agent's comparative judgements (which we take as primitive): whenever the agent is more confident in A then in B, A goes to the right of B. The impossible proposition $\bot$ will then be at the left end, the necessary proposition $\top$ at the right end. If the agent is equally confident in two propositions, they are stacked on top of each other at the same point on the line.

Now imagine we hold a ruler under this line in such a way that $\bot$ lies at 0 and $\top$ at 1. Every other proposition will then have a number between 0 and 1, given by its position along the line. If that's how we understand degrees of belief, to say that an agent's degree of belief in rain is 0.2 is to identify the relative position of rain in an agent's confidence ordering.

The Probability Representation Theorem tells us that if the confidence ordering satisfies the conditions I have described, then there will be exactly one way of assigning numbers to the propositions that respects the probability axioms: there is a unique probability measure Cr that *represents* the confidence ordering, meaning that $\text{Cr}(A) > \text{Cr}(B)$ whenever $A \succ B$, and $\text{Cr}(A) = \text{Cr}(B)$ whenever $A \sim B$. Assuming that an agent's degrees of belief satisfy the probability axioms therefore amounts to choosing a particular kind of ruler for measuring degrees of belief.

On this approach, any agent whose attitudes of comparative credence satisfy the six assumptions is guaranteed to have probabilistic credences, because the agent's credence function is *defined* as the unique probability measure Cr that represents her confidence ordering.

As you may imagine, this approach has also not gone unchallenged. One obvious question is whether we can take comparative confidence as primitive. If we can, a further question is whether the six assumptions are plausible as general norms of rationality. Transitivity, Normalization, and Boundedness look fairly safe, but the others have been questioned.

The missing sixth assumption is especially troublesome in this regard. As it turns out, the form of that assumption depends on whether the number of propositions ranked by $\succ$ is finite or infinite. In either case the condition is complicated – which makes it especially hard to treat it as a basic norm of rationality. Just to prove the point, here is the condition for the slightly simpler case of finitely many propositions:

The Sixth Assumption (finite version)

For any two sequences of propositions $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$ such that for every possible world w, the number of propositions in the first sequence that contain w equals the number of propositions in the second sequence that contain w, if $A_i \succsim B_i$ for all $i < n$, then $B_n \succsim A_n$.

3.7 Further reading

A thorough critique of Dutch Book arguments can be found in

- Alan Hájek: "Dutch Book Arguments" (2008).

For even more details and background information, have a look at the Stanford Encyclopedia entry

- Susan Vineberg: "Dutch Book Arguments" (2016).

If you're interested in the approach based on comparative credence, a good (though mathematically non-trivial) introduction is

- Peter Fishburn: "The Axioms of Subjective Probability" (1986).

Essay Question 3.1

Do you think the Dutch Book Theorems can teach us anything about epistemic rationality? If so, can you spell out how? If not, can you explain why not?

4 Further Constraints on Rational Belief

4.1 Belief and perception

We have looked at two assumptions about rational belief. The first, the MEU Principle, relates an agent's beliefs and desires to her choices. The second, probabilism, imposes an internal, structural constraint on rational beliefs: that they satisfy the axioms of probability. But there is more.

Example 4.1 (The Litmus Test)

You are unsure whether a certain liquid is acidic. Remembering that acid turns litmus paper red, you dip a piece of litmus paper into the liquid. The paper turns red.

When you see the paper turn red, your credence in the hypothesis that the liquid is acidic should increase. But as far as probabilism and the MEU Principle are concerned, you could just as well remain unsure whether the liquid is acidic or even become certain that it is *not* acidic, as long as your new credences are probabilistic and your choices maximize expected utility (by the light of your beliefs and desires).

So there are further norms on rational belief. In particular, there are rules for how beliefs should change in response to perceptual experience. Like the MEU Principle (and unlike probabilism), these rules state a connection between beliefs and something other than belief – perceptual experience. Informally speaking, the MEU Principle describes the causal "output" of beliefs: the effects an agent's beliefs (and desires) have on her behaviour. Now we turn to the "input" side. We want to know how a rational agent gets to have such-and-such beliefs in the first place.

To state a connection between perceptual experience and belief, we need a way to identify different kinds of perceptual experience. How do we do that? We could

identify perceptual experiences by their phenomenology, by "what it's like" to have the experience. But there is no canonical standard for expressing phenomenal qualities. Besides, we may want our norm to handle unconscious perceptions and the perceptions of artificial agents for whom it is doubtful whether they have any phenomenal experience.

An alternative to identifying perceptions by their phenomenology is to identify them by their physiology, by the neurochemical or electrical events that take place in the agent's sense organs. But that would go against the spirit of our general approach, which is to single out high-level patterns in rational cognition that are neutral on details of biological or electrical implementation.

The usual strategy is therefore to identify perceptions neither by their phenomenology nor by their physiology, but by their effect on rational belief. The idea is that perceptual experiences provide an agent with direct information about certain aspects of the world, so we can distinguish perceptual experiences by the information they provide. In the Litmus Test example, your visual experience tells you that the litmus paper has turned red. It does not directly tell you that the liquid is acidic; this is something you infer from the experience with the help of your background beliefs.

In the simplest and best known version of this model, we assume that the information conveyed to an agent by a given experience is captured by some proposition of which the agent becomes certain. The model can be extended to allow for cases in which the perceptual information is uncertain and equivocal, but we will stick to the simplest version.

4.2 Conditionalization

So assume through perceptual experience an agent learns some proposition E (for "evidence"), of which she rationally becomes certain. How should the rest of her beliefs change to take into account the new information?

Return to the Litmus Test. Let $\mathrm{Cr}_{\mathrm{old}}$ be your credence function before you dipped the paper into the liquid, and $\mathrm{Cr}_{\mathrm{new}}$ your credence function after seeing the paper turn red. If you are fairly confident that red litmus paper indicates acidity, you will also be confident, before dipping the paper, that the liquid is acidic *on the supposition that* the paper will turn red. So your initial degrees of belief might have been as follows.

$\text{Cr}_{\text{old}}(\textit{Acid}) = 1/2$.
$\text{Cr}_{\text{old}}(\textit{Acid}/\textit{Red}) = 9/10$.

What is your new credence in *Acid*, once you learn that the paper has turned red? Plausibly, it should be $9/10$. Your previous conditional credence in *Acid* given *Red* should turn into your new unconditional credence in *Acid*.

This kind of belief change is called **conditionalization**: we say that you conditionalized **on** the information *Red*. Let's formulate the general rule.

The Principle of Conditionalization

Upon receiving information E, a rational agent's new credence in any proposition A equals her previous credence in A conditional on E:

$$\text{Cr}_{\text{new}}(A) = \text{Cr}_{\text{old}}(A/E).$$

Here it is understood that the agent's perceptual experience leaves no room for doubts about E, and that E is the *total* information the agent acquires, rather than part of her new information. For example, if you see the paper turn red but at the same time notice the smell of ammonium hydroxide, which you know is alkaline, your credence in the *Acid* hypothesis may not increase to 0.9.

Exercise 4.1 ⋆

Assume $\text{Cr}_{\text{old}}(\textit{Snow}) = 0.3$, $\text{Cr}_{\text{old}}(\textit{Wind}) = 0.6$, and $\text{Cr}_{\text{old}}(\textit{Snow} \wedge \textit{Wind}) = 0.2$. By the Principle of Conditionalization, what is $\text{Cr}_{\text{new}}(\textit{Wind})$ if the agent finds out that it is snowing?

Exercise 4.2 ⋆⋆

Show from the definition of conditionalization and the rules of probability that if Cr_{new} results from Cr_{old} by conditionalizing on some information E with $\text{Cr}_{\text{old}}(E) > 0$, then $\text{Cr}_{\text{new}}(E) = 1$.

Exercise 4.3 ⋆⋆⋆

Assume that Cr_{new} results from Cr_{old} by conditionalizing on some information E with $\text{Cr}_{\text{old}}(E) > 0$, and that Cr_{old} satisfies the Kolmogorov axioms. Using the probability rules, show that Cr_{new} then also satisfies the Kolmogorov axioms. (You may use any of the derived rules from chapter 2. Hint for Kolmogorov's

axiom (ii): if A is logically necessary, then $A \wedge E$ is logically equivalent to E.)

The Principle of Conditionalization seems obvious enough. It is also supported by a range of arguments. For example, one can show that any agent who violates the Principle is vulnerable to a "diachronic Dutch Book" – a collection of bets, some offered before the arrival or the new information and some afterwards, that together amount to a sure loss. As you may have guessed, all these arguments are controversial. Let's skip them and instead look at some applications.

When computing $\mathrm{Cr}_{\mathrm{new}}(A)$, it is often helpful to expand $\mathrm{Cr}_{\mathrm{old}}(A/E)$ with the help of Bayes' Theorem. The Principle of Conditionalization then turns into the following (equivalent) norm, known as **Bayes' Rule**:

$$\mathrm{Cr}_{\mathrm{new}}(A) = \frac{\mathrm{Cr}_{\mathrm{old}}(E/A) \cdot \mathrm{Cr}_{\mathrm{old}}(A)}{\mathrm{Cr}_{\mathrm{old}}(E)}, \text{ provided } \mathrm{Cr}_{\mathrm{old}}(E) > 0.$$

The usefulness of this formulation comes from the fact that it is often much easier to evaluate the probability $\mathrm{Cr}_{\mathrm{old}}(E/A)$ of the evidence E conditional on some hypothesis A than to evaluate the probability $\mathrm{Cr}_{\mathrm{old}}(A/E)$ of the hypothesis A conditional on the evidence.

Let's do an example.

Example 4.2

2% of women in a certain population have breast cancer. A test is developed that correctly detects 95% of cancer cases but also gives a false positive result in 10% of cases without the cancer. A woman from the population takes the test, and gets a positive result. How confident should you be that the woman has breast cancer?

Let's imagine that you know all the statistical information before finding out about the test result. Knowing that the woman is from a population in which 2% of women have breast cancer, your initial credence in the hypothesis C that the woman has breast cancer should plausibly be 0.02. So $\mathrm{Cr}_{old}(C) = 0.02$. Moreover, since you know that the test yields a positive result in 95% of cancer cases, $\mathrm{Cr}_{\mathrm{old}}(P/C) = 0.95$, where P is the proposition that the test result is positive. Similarly, since the test yields a positive result in 10% of non-cancer cases, $\mathrm{Cr}_{\mathrm{old}}(P/\neg C) = 0.1$. Now we simply plug these numbers into Bayes' Rule,

expanding the denominator by the Law of Total Probability:

$$\begin{aligned}\mathrm{Cr}_{\mathrm{new}}(C) &= \frac{\mathrm{Cr}_{\mathrm{old}}(P/C)\cdot\mathrm{Cr}_{\mathrm{old}}(C)}{\mathrm{Cr}_{\mathrm{old}}(P/C)\cdot\mathrm{Cr}_{\mathrm{old}}(C)+\mathrm{Cr}_{\mathrm{old}}(P/\neg C)\cdot\mathrm{Cr}_{\mathrm{old}}(\neg C)}\\ &= \frac{0.95\cdot 0.02}{0.95\cdot 0.02+0.1\cdot 0.99} = \frac{0.019}{0.019+0.098} = 0.16.\end{aligned}$$

The answer is much lower than many people think – including many trained physicians. But it makes intuitive sense. Imagine we took a large sample of 1000 women from the population. We would expect around 2%, or 20 women, in the sample to have breast cancer. If we tested all women in the sample, we would expect around 95% of those with cancer to test positive. That's 95% of $20 = 19$ women. Of the 980 women without cancer, we would expect around 10%, or 98 women, to test positive. The total number of women who would test positive would therefore be $19 + 98 = 117$. Of these 117, 19 actually have cancer. So the chance that a woman who tests positive has cancer is $19/117 = 0.16$. If you look back at the above application of Bayes' Theorem, you can see that it basically encodes this line of reasoning.

The tendency to overestimate the significance of tests in cases like example 4.2 is known as the **base rate fallacy** because it is assumed to arise from neglecting the low "base rate" of 2%.

Exercise 4.4 ⋆⋆

Box A contains two black balls. Box B contains one black ball and one white ball. I choose a box at random and blindly draw a ball. The ball is black. How confident should you be that I chose box A? (Explain briefly.)

Exercise 4.5 (The Prosecutor's Fallacy) ⋆⋆⋆

A murder has been committed on a remote island with a million inhabitants. In a database of blood donors, detectives find a record whose DNA seems to match the perpetrator's DNA from the crime scene. The DNA test is very reliable: the probability that it finds a match between distinct people is 1 in 100,000. So the person with the matching DNA is arrested and brought to court. The prosecutor argues that the probability that the defendant is innocent is 1/100,000. Is that correct? As a member of the jury, how confident should you be in the defendant's guilt?

4.3 The Principle of Indifference

If an agent's beliefs evolve by conditionalization, can we be sure that her beliefs will adequately reflect all the evidence she receives over time? No. If the agent starts out with crazy beliefs, conditionalization will not make her sane.

Example 4.3

You are stranded on a remote island, which you find inhabited by a strange kind of flightless bird. In the first ten days of your stay on the island, you see 100 birds, all of which are green.

Plausibly, you should be fairly confident that the 101st bird will also be green. The Principle of Conditionalization does not ensure this. To see why, let H be the proposition that the first 100 birds you encounter on the island are atypical in colour. Suppose when you first arrived on the island you were convinced of H – for no good reason. The observation of 100 green birds does not challenge that conviction, so after conditionalizing on these observations you are still confident in H. And if you believe that the first 100 birds you encountered were green and also atypical in colour, then you'll expect the 101st bird to have a different colour. So if we think you should be confident that the 101st bird will be green, we have to say that you should not be confident in H before receiving any relevant evidence.

What we see here is Hume's problem of induction. As Hume pointed out, there is no logical guarantee that the future will resemble the past, or that the unobserved parts of the world resemble the observed. The colour of the 101st bird is not entailed by the colour of the first 100 birds. To infer that the 101st bird is green we thus need a further premise about the "uniformity of nature": that the 101st bird is likely to have the same colour as the first 100 birds. How do we know this? We may have inferred it from our experiences at 100 other islands, but to conclude that the lessons from these islands carry over to the present island, we need another premise about the uniformity of nature. Ultimately, some such premise must be taken for granted.

In Bayesian terms, this means that we have to impose constraints on what an agent may believe *without any relevant evidence*. Scientifically minded people sometimes feel uneasy about such constraints, and therefore speak about the **problem of the priors**. An agent's **priors** (or "ultimate priors") are her credences before receiving any evidence. The problem of the priors is to explain what

rational priors should look like.

So what should you believe if you have no evidence at all about a certain subject matter? A natural thought is that you should be maximally open-minded. For example, if you know that one of three people has committed a murder, but you have no further information about the case, then you should give equal credence to the three possibilities. More generally, the following principle looks appealing.

The Principle of Indifference

If $A_1, \ldots, A_n$ are n propositions exactly one of which must be true, and an agent has no evidence relevant to these propositions, then her credence in each of the propositions should be $1/n$.

Unfortunately, the Principle of Indifference can't be right, because it is inconsistent. For example, suppose you have no information about the colour of my hat. Here are two possibilities:

R: The hat is red.
$\neg R$: The hat is not red.

Exactly one of these must be true. By the Principle of Indifference, you should therefore give credence $1/2$ to both R and $\neg R$. But we can also divide $\neg R$ into several possibilities:

R: The hat is red.
B: The hat is blue.
G: The hat is green.
Y: The hat is yellow.
O: The hat has some other colour.

Again exactly one of these must be true, so by the Principle of Indifference, you should give credence $1/5$ to each. So the Principle entails that your credence in R should be $1/2$ and also that it should be $1/5$!

Some have concluded that in cases like these, rationality really does require you to have *several* credence functions: relative to one of your credence functions, R has probability $1/2$, relative to another, it has probability $1/5$. I'll set this view aside for now, but we will return to it in section 11.5.

Another response is to say more about the propositions $A_1, \ldots, A_n$ to which the Principle applies. Intuitively, you might say, the Principle does not hold for R

and $\neg R$ because these two propositions are not on a par: there are more ways of being non-red than there are of being red. Unfortunately, it is hard to make this intuition precise, and harder still to turn it into a general rule, as the following exercise illustrates.

Exercise 4.6 ⋆⋆

A cube is hidden in a box. A sticker on the box reveals that the cube has a side length of at least 2 cm, but less than 4 cm. So here are two possibilities:

S: The cube's side length lies between 2 cm and 3 cm (excluding 3).
L: The cube's side length lies between 3 cm and 4 cm (excluding 4).

The intervals have the same length, so S and L are intuitively on a par. We might infer that you should give credence $1/2$ to both S and L. But now observe that if a cube has side length x, then the cube's volume is x^3.

(a) Can you restate the propositions S and L in terms of volume?
(b) What should your credence in S be if you treat equally sized ranges of volume as equally likely?

Another problem with the Principle of Indifference is that it actually clashes with the "uniformity of nature" assumption required for inductive inference. Return to example 4.3. For simplicity, let's assume you know in advance that any bird on the island can only be green or red. So there are four possibilities regarding the first two birds you might see:

GG: Both birds are green.
GR: The first bird is green, the second is red.
RG: The first bird is red, the second is green.
RR: Both birds are red.

By the Principle of Indifference, you should give credence $1/4$ to each of these possibilities. (They are also intuitively on a par.) Now what happens when you see the first bird, which is green? Your evidence E rules out RG and RR. By the Principle of Conditionalization, your new credence in GG equals your previous credence in GG conditional on E, which (as you should check) is $1/2$. So after having seen the first green bird, your credence in the hypothesis that the next bird will be green is $1/2$. By the same reasoning, your credence in the hypothesis that the third bird will be green after having seen two green birds, is also $1/2$.

In general, no matter how many green birds you see, your credence in the next bird being green will remain at $1/2$. In other words, the Principle of Indifference makes it impossible to draw inductive inferences from experience.

Despite these problems, many instances of the Principle of Indifference look very plausible. If you're investigating a murder, and judge that suspect A is three times as likely as suspect B to be the murderer, then there had better be some reason for this judgement. In the absence of any relevant evidence, your judgement would be irrational. Several authors have attempted to turn examples like this into a fully general principle which, unlike the classical Principle of Indifference, is consistent and allows for inductive inference, but there is no agreement on whether this can be done and on what the resulting principle should look like. In this form, at least, the "problem of the priors" remains open.

4.4 Probability coordination

We turn from the highly controversial Principle of Indifference to another norm for rational priors that is almost universally accepted among Bayesians. The norm connects subjective probability with objective probability, and plays a central role in Bayesian confirmation theory and Bayesian statistics.

The Probability Coordination Principle

If an agent has no evidence about some proposition A, then her credence in A on the supposition that the objective probability of A is x, should be x:

$$\mathrm{Cr}_0(A/\mathrm{Pr}(A)=x) = x$$

Here 'Pr' stands for any kind of objective probability, such as relative frequency or quantum physical chance. I've added a subscript '0' to 'Cr' to indicate that the agent in question has no evidence relevant to A. With 'Pr' understood as quantum physical chance, the Probability Coordination Principle is also known as the 'Principal Principle'.

We have unwittingly assumed the Probability Coordination Principle all along. In example 4.2, for instance, we assumed that if all you know about a woman is that she is from a population in which 2% of women have breast cancer, then your credence in the hypothesis that she has breast cancer should be 0.02. This is clearly not entailed by the Kolmogorov axioms. It is, however, entailed by the Probability Coordination Principle, assuming that your credence can be modelled

as resulting from a prior state, in which you had no evidence at all about the woman, by conditionalizing on the statistical information that the cancer rate is 2%.

To see why the Probability Coordination Principle is stated in terms of conditional credence, consider a typical case of testing scientific hypotheses. Often such hypotheses only make statistical predictions: they entail that under circumstances C, there is a probability of x that outcome O will occur.

Concretely, suppose you are undecided between two theories, H_1 and H_2, giving credence $1/2$ to each. H_1 says that under circumstances C, the probability of O is 0.9; H_2 says it is 0.3. You set up an experiment with circumstances C and observe outcome O. How does that affect your credence in H_1 and H_2? By Bayes' Rule,

$$\mathrm{Cr}_{\mathrm{new}}(H_1) = \frac{\mathrm{Cr}_{\mathrm{old}}(O/H_1) \cdot \mathrm{Cr}_{\mathrm{old}}(H_1)}{\mathrm{Cr}_{\mathrm{old}}(O/H_1) \cdot \mathrm{Cr}_{\mathrm{old}}(H_1) + \mathrm{Cr}_{\mathrm{old}}(O/H_2) \cdot \mathrm{Cr}_{\mathrm{old}}(H_2)}.$$

We know that $\mathrm{Cr}_{\mathrm{old}}(H_1) = \mathrm{Cr}_{\mathrm{old}}(H_2) = 1/2$. The Probability Coordination Principle tells us that $\mathrm{Cr}_{\mathrm{old}}(O/H_1) = 0.9$ and $\mathrm{Cr}_{\mathrm{old}}(O/H_2) = 0.3$. Thus

$$\mathrm{Cr}_{\mathrm{new}}(H_1) = \frac{0.9 \cdot 0.5}{0.9 \cdot 0.5 + 0.3 \cdot 0.5} = 0.75.$$

So your credence in H_1 should increase to $3/4$, and your credence in H_2 should decrease to $1/4$.

Exercise 4.7 ⋆⋆

You are unsure whether a certain coin is biased 2:1 towards heads or 2:1 towards tails; initially you give credence $1/2$ to each possibility. Then you toss the coin twice, and both times it comes up heads. What is your new credence concerning the coin's bias? (If a coin is biased 2:1 towards heads, then heads has an objective probability of $2/3$.)

4.5 Anthropic reasoning

How confident should an ideal agent, without any relevant evidence, be that she exists? A strange question, but a question that sometimes comes up in cosmology and certain philosophical puzzles.

The following puzzle is due to Nick Bostrom.

Example 4.4 (God's Coin Toss)

At the beginning of time, God flips a fair coin. If the coin lands heads, she creates two people in two rooms, one with blue eyes and one with green eyes. If the coin lands tails, God creates only one room with a blue-eyed person in it. You wake up, and God informs you of these facts. Then you look in the mirror and see that your eyes are blue. How confident should you be that God's coin landed heads?

At first, you might think the answer is $1/2$ on the grounds that the objective probability of heads is $1/2$ and your evidence of having blue eyes is equally compatible with heads and tails. But notice that if you had found your eyes to be green, then you could have inferred with certainty that God's coin landed heads. And if some evidence E increases the probability of a hypothesis H, then $\neg E$ must decrease the probability of H. So finding your eyes to be blue should decrease your credence in heads. More specifically, if $\mathrm{Cr}_{\mathrm{old}}(\mathit{Heads}) = 1/2$, then by Bayes' Rule, $\mathrm{Cr}_{\mathrm{new}}(\mathit{Heads}) = 1/3$.

Exercise 4.8 ⋆

Show this. That is, assume $\mathrm{Cr}_{\mathrm{old}}(\mathit{Heads}) = 1/2$, and use Bayes' Rule to derive that $\mathrm{Cr}_{\mathrm{new}}(\mathit{Heads}) = 1/3$.

To conclude that your credence in *Heads* should be $1/3$, we would have to assume that $\mathrm{Cr}_{\mathrm{old}}(\mathit{Heads}) = 1/2$. But that, too, could be questioned. To be sure, the Probability Coordination Principle requires that $\mathrm{Cr}_{\mathrm{old}}(\mathit{Heads}) = 1/2$ *if you have no other relevant evidence*. But one might argue that you do have further relevant evidence – namely, the evidence that you exist.

Why should that be relevant? The idea is that the more people there are in a possible world, the more likely it is that one of these people is you. In a heads world, there are two people; so the chance that you are one of them is twice the chance that you're the single person in a tails world. By that line of thought, the observation that you exist, which you make before looking in the mirror, should increase your credence in heads from $1/2$ to $2/3$. Finding that your eyes are blue then reduces it back to $1/2$.

Exercise 4.9 ⋆⋆

Show that

(a) if $\mathrm{Cr}(\mathit{Heads}) = 1/2$ and $\mathrm{Cr}(\mathit{Exist}/\mathit{Heads}) = 2 \cdot \mathrm{Cr}(\mathit{Exist}/\mathit{Tails})$, then by Bayes' Theorem, $\mathrm{Cr}(\mathit{Heads}/\mathit{Exist}) = 2/3$;
(b) Assuming $\mathrm{Cr}_{\mathrm{old}}(\mathit{Heads}) = 2/3$, then after seeing that your eyes are blue, $\mathrm{Cr}_{\mathrm{new}}(\mathit{Heads}) = 1/2$.

If you are sceptical about this argument for $\mathrm{Cr}_{\mathrm{new}}(\mathit{Heads}) = 1/2$, you are not alone. Among other things, the argument seems to assume that rational agents should initially give significant credence to the hypothesis that they don't exist, and it's not clear why that should be a requirement of rationality.

Anyway, let's give a name to the problematic assumption:

Dubious Principle

If H_1 is the hypothesis that there are n people in total, and H_2 says that there are k people, then $\mathrm{Cr}_0(\mathit{Exist}/H_1) = \mathrm{Cr}_0(\mathit{Exist}/H_2) \cdot n/k$, where Cr_0 is the credence function of a rational agent without any evidence, and *Exist* is the proposition that the agent exists.

So far, all this may look like idle sophistry. But now consider the hypothesis that the human race will go extinct within the next few years, at a point where the total number of people who ever lived will be around 100 billion. By contrast, if humankind continues to prosper for another million years or so, the total number of people who ever lived will be at least a thousand times greater. To keep the maths easy, let's pretend that these are the only two possibilities. Call the first *Doom* and the second *No Doom*. A priori – in the absence of any evidence – you might think *Doom* and *No Doom* deserve roughly equal credence. But here's a piece of evidence you have (call it *Early*): you are one of the first 100 billion people. And this dramatically increases the probability of *Doom*.

Let's crunch the numbers. On the supposition that the total number of people is 100 trillion (100,000 billion), only $1/1000$ of all people are among the first 100 billion. So the prior probability that you are one of the first 100 billion is arguably $1/1000$. By contrast, on the supposition that the total number of people is 100 billion, the probability that you are one of the first 100 billion is 1. So, by Bayes'

Theorem,

$$\mathrm{Cr}(\mathit{Doom}/\mathit{Early}) = \frac{\mathrm{Cr}(\mathit{Early}/\mathit{Doom}) \cdot \mathrm{Cr}(\mathit{Doom})}{\mathrm{Cr}(\mathit{Early}/\mathit{Doom}) \cdot \mathrm{Cr}(\mathit{Doom}) + \mathrm{Cr}(\mathit{Early}/\neg\mathit{Doom}) \cdot \mathrm{Cr}(\neg\mathit{Doom})}$$

$$= \frac{1 \cdot \mathrm{Cr}(\mathit{Doom})}{1 \cdot \mathrm{Cr}(\mathit{Doom}) + 0.001 \cdot \mathrm{Cr}(\neg\mathit{Doom})}$$

If $\mathrm{Cr}_0(\mathit{Doom}) = {}^1/_2$, it follows that $\mathrm{Cr}_0(\mathit{Doom}/\mathit{Early}) = 1000/1001 \approx 0.999$. Taking into account the fact that you are among the first 100 billion people who ever lived, it seems that you should be almost certain that humankind is about to go extinct!

The argument we've just rehearsed is known as the **doomsday argument**. The conclusion becomes a little less striking if we take into account that there are more possibilities than *Doom* and *No Doom*, but the upshot remains the same: we should be highly confident that we are among the last people to have lived.

You may not be surprised to hear that. After all, we face a long list of existential threats – nuclear war, global warming, pandemics, hostile AI, and so on. What's surprising is that none of these threats are taken into account in the doomsday argument. The conclusion is reached solely on the basis of population statistics.

Given that *Early* strongly increases the probability of *Doom*, to block the conclusion that *Doom* is highly probable, we would have to assume that the prior probability of *Doom*, before taking into account *Early*, is much lower than the prior probability of *No Doom*. But why should that be so? A priori, the hypothesis that the total number of people is 100 billion is roughly on a par with the hypothesis that the number is 100,000 billion: taking the second to be 1000 times more likely than the first, without any relevant evidence, surely seems irrational.

But perhaps we have evidence that favours *No Doom* over *Doom* – namely, the evidence of our existence. By the Dubious Principle, $\mathrm{Cr}_0(\mathit{Exist}/\neg\mathit{Doom})$ is 1000 times greater than $\mathrm{Cr}_0(\mathit{Exist}/\mathit{Doom})$. Consequently, if $\mathrm{Cr}_0(\mathit{Doom}) = {}^1/_2$, then $\mathrm{Cr}_0(\mathit{Doom}/\mathit{Exist}) = {}^1/_{1000}$ and $\mathrm{Cr}_0(\mathit{Doom}/\mathit{Exist} \wedge \mathit{Early}) = {}^1/_2$.

So it would be comforting if the Dubious Principle were true after all.

4.6 Further reading

Chapter 15 of

- Ian Hacking: *An Introduction to Probability and Inductive Logic* (2001)

goes into some more details about conditionalization.

Several philosophers have recently defended versions of the Principle of Indifference, and the more general claim that there is a unique rational prior credence function. A critical overview and discussion can be found in

- Christopher G. Meacham: "Impermissive Bayesianism" (2014).

For more on the Doomsday Argument and related puzzles, see

- Nick Bostrom: "The Doomsday Argument, Adam & Eve, UN++, and Quantum Joe" (2001).

(What I call the 'Dubious Principle' is Bostrom's 'Self-Indication Assumption', SIA.)

Essay Question 4.1

Can you think of another way to block the Doomsday Argument, without relying on the Dubious Principle?

5 Utility

5.1 Two conceptions of utility

Daniel Bernoulli realized that rational agents don't always maximize expected monetary payoff: £1000 has more utility for a pauper than for a rich man. But what is utility?

Until the early 20th century, utility was widely understood to be some kind of psychological quantity, often identified with pleasure or absence of pain. On that account, an outcome has high utility for an agent to the extent that it increases the agent's pleasure and/or decreases her pain.

Let's assume for the sake of the argument that one can represent an agent's total amount of pleasure and pain by a single number – the agent's 'degree of pleasure'. Can we understand utility as degree of pleasure? The answer depends on what role we want the concept of utility to play.

One such role lies in ethics. Here, **utilitarianism** is the view that an act is morally right just in case it would bring about greater total utility, for all people combined, than any other available act. In the context of utilitarianism, identifying utility with degree of pleasure implies that only pleasure and pain have intrinsic moral value; everything else – autonomy, integrity, respect of human rights, and so on – would be morally relevant only insofar as it causes pleasure or pain. This assumption is known as **ethical hedonism**. We will not pursue it any further.

Exercise 5.1 ⋆

Suppose that money has declining marginal utility, and that the utility of money is the same for all people (so a net wealth of £1000 is as good for me as it is for you). Without any further assumptions about utility, it follows that if one person has more money than another, then their combined utility would increase if the wealthier person gave some of her money to the poorer person, decreasing the gap in wealth. Explain why!

Another role for the concept of utility lies in the theory of practical rationality. Here the MEU Principle states that (practically) rational agents choose acts which maximize the probability-weighted average of the utility of any possible outcome. If we identify utility with degree of pleasure, the MEU principle turns into what we might call the 'MEP Principle':

The MEP Principle

Rational agents maximize their expected degree of pleasure.

An act's *expected degree of pleasure* is the probability-weighted average of the degree of pleasure that might result from the act.

The MEP Principle is a form of **psychological hedonism**. Psychological hedonism is the view that the only thing that ultimately motivates people is their own pleasure and pain.

The founding fathers of modern utilitarianism, Jeremy Bentham and John Stuart Mill, had sympathies for both ethical hedonism and psychological hedonism, and so the two conceptions of utility – the two roles associated with the word 'utility' – were not properly distinguished. Today, both kinds of hedonism have long fallen out of fashion, but the two conceptions are still often conflated.

For the most part, contemporary utilitarians hold that the standard of moral rightness is the total *welfare* or *well-being* produced by an act, which is not assumed to coincide with total degree of pleasure. Thus 'utility' is nowadays often used as a synonym for 'welfare' or 'well-being'. But the word is also widely used in the other sense, to denote whatever motivates (rational) agents.

Some have argued that the two uses actually coincide: that the only thing that motivates rational agents is their own welfare or well-being. This may or may not be true. But it needs to be backed up by data and argument; it does not become true through sloppy use of language.

In these notes, 'utility' is only used in the second sense, which fits the official definition in economics textbooks. That is, the utility of a state of affairs measures the extent to which the agent in question wants the state to obtain. We do not assume that the only thing agents ultimately want is to increase their degree of pleasure, their welfare, their well-being, or anything like that.

Note that psychological hedonism, or the slightly more liberal claim that people only care about their own welfare, is at most a contingent fact about humans. One can easily imagine agents who are motivated by other things. There is no contradiction in the hypothesis that an agent chooses acts that she knows will

decrease her pleasure or welfare – a mother who knowingly takes on hardships for the benefit of her children, a soldier who intentionally chooses a painful death in order to save her comrades, and so on. Psychological hedonists claim that humans would never consciously do such things: whenever an agent sacrifices her own good to benefit others, she mistakenly believes that her choice will actually make herself better off than the alternatives. Again, we don't need to argue over whether this is true. The important point is that utility, as we use the term, does not *mean* the same as degree of pleasure or welfare or well-being.

A hedonist might object that while it is conceivable that an agent is motivated by other things than her personal pleasure, such agents would be irrational. After all, the MEP Principle only states that *rational* agents maximize their expected degree of pleasure; it doesn't cover irrational agents.

This brings us to a tricky issue: what do we mean by 'rational'? The label 'rational' is sometimes associated with cold-hearted selfishness. A rational agent is assumed to be an agent who always looks out for her own advantage, with no concern for others. This idea of "economic rationality" has its use, but it is not our topic. The kind of rationality we're interested in is a more abstract and minimal notion. Intuitively, it is the idea of "making sense". If you want to reduce animal suffering, and you know you can achieve this by eating less meat, then it makes sense that you eat less meat. If you are sure that a picnic will be cancelled if it rains, and you learn that it rains, then it doesn't make sense to believe that the picnic will go ahead. The model we are studying is a model of agents who "make sense" in this kind of way.

Even if we were interested in the cold-hearted and selfish sense of rationality, we shouldn't *define* utility as degree of pleasure or welfare. Consider a hypothetical agent who cares not just about herself, and who sacrifices some of her own good to reduce the pain of others. The agent is "irrational" in the cold-hearted and selfish sense. But what is irrational about her? Does the fault lie in her beliefs, in her goals, or in the way she brings these together to make choices? Plausibly, the "fault" lies in her goals. Her concern for others is what goes against the standards of cold-hearted and selfish rationality. However, if we were to define utility as degree of pleasure, we would have to say that the agent violates the basic norm of practical rationality, the MEU Principle, which states how credences and utilities combine to determine action.

The point generalizes. Imagine a person in an abusive relationship who is manipulated into doing things that hurt her. We might reasonably think that the person shouldn't do these things; it is against her interest to do them. But what

is at fault? Arguably, the fault lies in her (manipulated) desires. What the person does may well be in line with what she wants to achieve – in particular, with her strong desire to please her partner. But a healthy, self-respecting person, we think, should have other desires.

By understanding utility as a measure of whatever the agent in question desires, we do not automatically treat these desires as rational. Our usage of 'utility' allows us to say that the person in the abusive relationship shouldn't do what she is doing, because she should have different desires that would not support her actions.

5.2 Sources of utility

On our usage, an outcome's utility measures the extent to which the agent desires the outcome. But the word 'desire' can be misleading, and has given rise to further misunderstandings. For one thing, we also need to cover "negative desire". Being hungry might have greater utility for you than being dead, even though you do not desire either. More importantly, the word 'desire' is often associated with a particular type of motivational state: with unreflective urges and cravings, in contrast to more considerate and dignified motives. I might say that I got up early in the morning despite my strong desire to stay in bed; I got up not because I desired to get up, but because I had to. On this usage, my desires contrast with my sense of duty.

Utility is meant to comprise everything that motivates the agent, all the reasons she has for and against a particular action. As such, utility is an umbrella term for a diverse set of psychological states or events. We can be motivated by cravings and appetites, by moral commitments, by our image of the kind of person we want to be, by an overwhelming feeling of terror or love, and so on. These factors may not be conscious: there is good evidence that our true motives are often not what we think or say. An agent's utility function represents her true motives, and all of them.

Why should we believe that all the factors that motivate an agent can be amalgamated into a single numerical quantity? Would it not be better to allow for a whole range of utility functions: moral utility, emotional utility, and so on? We could certainly do that. But there are reasons to think that there must also be an amalgamated, all-things-considered utility (although the determinacy and numerical precision of utility functions is obviously an idealisation). In particular,

when you face a decision, you have to make a single choice. You can't choose one act on moral grounds and a different act on emotional grounds. Somehow, all your motives and reasons have to be weighed against each other to arrive at an overall ranking of your options.

We will have a brief look at the weighing of different considerations in chapter 7, but to a large extent this is really a topic for empirical psychology and neuroscience. If it turns out that there are 23 distinct factors that influence our motivation in an intricate network of inhibitation and reinforcement, then so be it. We will model the whole network by a single utility function, in part because we want to stay neutral on "lower-level" details that can vary from agent to agent. But it's important to keep in mind that a lot of interesting and complicated psychology is hiding under the seemingly simple concept of subjective utility.

Above I mentioned that our use of 'utility' fits the official definition in economics textbooks. (The textbooks actually define 'utility' in terms of 'preference'; we'll come to that in the next chapter. It doesn't affect the present point.) In practice, however, economists and other social scientists often ignore most sources of human motivation and fall back onto a naive interpretation of utility in terms of material wealth.

Consider the following example.

Example 5.1 (The endowment effect)

Emily's favourite band is playing in town. The tickets cost £70, and Emily is not willing to pay that much. Emily's neighbour Fred has a ticket but can't go to the concert, so he sells his ticket to Emily for £50. The day before the concert, all the tickets are sold out, and another of Emily's neighbours, George, asks Emily if she would sell her ticket to him for £70. Emily declines.

The kind of behaviour displayed by Emily is quite common. It is also widely claimed to contradict expected utility theory. The idea is that if Emily isn't willing to sell the ticket for £70, then having the ticket is worth more than having £70 to her, so she should have bought the ticket for £70 at the outset.

But it is not hard to understand what happened. By the time George approaches Emily, she has made plans for going to the concert; she is looking forward to the event. In that context, giving up the ticket would be a serious disappointment; it would also mean that she has to make new plans for tomorrow evening. In short, for Emily, *giving up a ticket she previously owned* is worse than *never owning the ticket in the first place*. Given these values, her behaviour is perfectly in line with

the MEU Principle.

Exercise 5.2 ⋆⋆

Draw an adequate decision matrix for Emily's decision problem when she first considered buying the ticket, and another matrix for her decision problem when she was approached by George. (There is no relevant uncertainty, so the matrices have only one state.)

So Emily's behaviour does not violate the MEU Principle. Indeed, no pattern of behaviour whatsoever can, all by itself, violate the MEU Principle. After all, for any pattern of behaviour, we can imagine that the agent has a basic desire to display just that pattern of behaviour. Displaying the behaviour then evidently maximizes expected utility.

So if we are interested in the MEU Principle as a descriptive hypothesis about real people's choices, and we interpret 'utility' to measure whatever people actually care about, then the Principle is, in a sense, unfalsifiable. Whenever an agent seems to violate the MEU Principle, we can postulate beliefs and desires that would make her choices conform to the principle. Social scientists sometimes point at this fact in support of their decision to re-interpret 'utility' as a function of material goods. A scientific hypothesis, they assume, is only worth taking seriously if it can turn out to be false. But a lot of respectable scientific theories are unfalsifiable *in isolation*. Philosophers of science have long realized that one can generally only test scientific hypotheses in conjunction with a whole range of background assumptions.

The same is true for the MEU Principle, understood as a descriptive hypothesis. *Given* some assumptions about an agent's beliefs and desires, we can easily find that her choices do not conform to the MEU Principle. And often we do have pretty good evidence about the relevant beliefs and desires. For example, it is fairly safe to assume that participants in the world chess tournament want to win their games, and that they are aware of the current position of the pieces in the game.

All that said, the model we are studying also has applications in which the utilities do not represent the agent's desires, or in which the credences do not represent her beliefs. For example, the board of directors of a company may want to know what corporate decisions would ideally promote shareholder value in the light of such-and-such information. Here the utility function could be derived from the (stipulated) goal of maximizing shareholder value, and the

credence function could be derived from the information at hand. Neither of these needs to match the beliefs and desires of any individual member of the board. Similarly, in ethics, the question may arise what an agent ought to do, from a moral perspective, in a case where she is not sure whether a given choice is right or wrong. Here the relevant utility function might be derived from an ethical system (utilitarianism, perhaps) and again need not match the agent's actual desires.

Exercise 5.3 ⋆⋆

Some choices can predictably change our desires. One might argue that a rational agent should be guided not by her present desires, but by the desires she will have as a result of the choice.

For example, suppose you can decide right now how many drinks you will have tonight: zero, one, or two. If you're sober, you prefer to have one drink rather than zero or two. But if you have a drink, you sometimes prefer to have another. Draw a matrix for your decision problem assuming that your goal is to maximize your expected future utility.

5.3 The structure of utility

Now that we know what utility is, let's have a closer look at its formal structure.

First of all, what are the bearers of utility? In ordinary language, we often say that people desire *things*: tea, cake, a concert ticket, a larger flat. This fits the economist tradition of identifying the bearers of utility with material goods, or "commodity bundles". But if we want to allow for the entire range of possible desires, we need a broader conception. Perhaps you desire that your friends are happy, that it won't rain tomorrow, that so-and-so will win the next elections. Here the object of desire isn't a particular thing, but a possible state of the world. In fact, even when we say that people desire things, plausibly the desire is really directed at a possible state of the world. When you desire tea, you desire to *drink the tea*. Your desire wouldn't be satisfied if I gave you a certificate of ownership for a cup of tea that is locked away in a safe.

So we'll assume that the objects of desire are the same kinds of things as the objects of belief: propositions, or possible states of the world. We will also assume that *typically*, whenever an agent assigns utility to some propositions, then she also assigns utility to arbitrary negations, conjunctions, and disjunctions of these

propositions. But we will have to make a few exceptions. In particular, on the way we will understand utility, logically impossible propositions ($A \wedge \neg A$) cannot have a well-defined utility. The reason for this will become clear soon.

First, let's investigate how an agent's desires towards logically related propositions are related to one another. For example, suppose you assign high utility to the proposition that it won't rain tomorrow (perhaps because you want to go on a picnic). Then you should plausibly assign low utility to the proposition that it *will* rain. It doesn't make sense to hope that it will rain and also that it won't rain. In this respect, desire resembles belief: if you are confident that it will rain, you can't also be confident that it won't rain. The Negation Rule of probability captures the exact relationship between $\mathrm{Cr}(A)$ and $\mathrm{Cr}(\neg A)$, stating that $\mathrm{Cr}(\neg A) = 1 - \mathrm{Cr}(A)$. Does the rule also hold for utility? More generally, do utilities satisfy the Kolmogorov axioms? It will be instructive to go through the three axioms.

Kolmogorov's axiom (i) states that probabilities range from 0 to 1. If there are upper and lower bounds on utility, we could adopt axiom (i) for utilities as a convention of measurement: we simply use 1 for the upper bound and 0 for the lower bound. However, it is not obvious that there are such bounds. Couldn't there be an infinite series $A_1, A_2, A_3, \ldots$ of propositions of increasing utility in which the difference in utility between successive propositions is always the same? If there is such a series, then utility can't be measured by numbers between 0 and 1. Philosophers are divided over the question. Some think utility must be **bounded**, others think it can be unbounded. There are arguments for either side, but we will not pause to look at them.

Kolmogorov's axiom (ii) states that logically necessary propositions have probability 1. If utilities satisfy the probability axioms, this would mean that logically necessary propositions have maximal utility. However much you desire that it won't rain tomorrow, your desire that *it either will or won't rain* should be at least as great.

This does not look plausible. Intuitively, if something is certain to be the case, it makes no sense to desire it. But this could mean two things. It could mean that degrees of desire are not even defined for logically necessary propositions. Or it could mean that an agent should always be indifferent towards logically necessary propositions – neither wanting them to be the case nor wanting them to not be the case. Our common-sense conception of desire arguably sides with the first option: if you are certain of something, even asking how strongly you desire it seems ill-posed. But the issue isn't clear. For our purposes, it proves

more convenient to go with the second option. So we will say that even logically necessary propositions have well-defined utility, but we will treat their utility as "neutral".

Axiom (iii) states that if A and B are logically incompatible, then the probability of $A \vee B$ equals the sum of the probabilities of A and B. To illustrate, suppose there are three possible locations for a picnic: Alder Park, Buckeye Park, and Cedar Park. Alder Park and Buckeye Park would be convenient for you; Cedar Park would not. Now how much do you desire that the picnic takes place in *either Alder Park or Buckeye Park*? If axiom (iii) holds for utilities, then if you desire Alder Park and Buckeye Park to equal degree x, then your utility for the disjunction should be $2x$: you should be more pleased to learn that the picnic takes place in either Alder Park or Buckeye Park than to learn that it takes place in Alder Park. That clearly doesn't make sense. So axiom (iii) also fails.

What is the true connection between the utility of $A \vee B$ and the utilities of A and B? Intuitively, if A and B have equal utility x, then the utility of $A \vee B$ should also be x. What if the utilities of A and B are not equal? What if, say, $U(A) > U(B)$? Then the utility of $A \vee B$ should plausibly lie in between the utilities of A and B:

$$U(A) \geq U(A \vee B) \geq U(B).$$

That is, if Alder Park is your first preference and Buckeye your second, then the disjunction *either Alder Park or Buckeye Park* can't be worse than Buckeye Park or better than Alder Park. But where does $U(A \vee B)$ lie in between $U(A)$ and $U(B)$? At the mid-point?

Suppose you prefer Alder Park to Buckeye Park, and Buckeye Park to Cedar Park. You think it is highly unlikely that the picnic will take place in Buckeye Park. Now how pleased would you be to learn the picnic won't take place in Cedar Park – equivalently, that it will take place either in Alder Park or in Buckeye Park? You should be quite pleased. If you're confident that B is false, then $U(A \vee B)$ should plausibly be close to $U(A)$. If you're confident that A is false, then $U(A \vee B)$ should be near $U(B)$.

So your utilities depend on your beliefs! On reflection, this should not come as a surprise. A lot of the things we desire we only desire because we have certain beliefs. If you want to buy a hammer to hang up a picture, then your desire for the hammer is based (in part) on your belief that the hammer will allow you to hang up the picture.

So here is the general rule for $U(A \vee B)$, assuming A and B are incompatible.

The rule was discovered by Richard Jeffrey in the 1960s and is the *only* basic rule of utility.

Jeffrey's Axiom

If A and B are logically incompatible, then

$$U(A \vee B) = U(A) \cdot \mathrm{Cr}(A/(A \vee B)) + U(B) \cdot \mathrm{Cr}(B/(A \vee B)).$$

In words: the utility of $A \vee B$ is the weighted average of the utility of A and the utility of B, weighted by the probability of the two disjuncts, conditional on $A \vee B$.

Why 'conditional on $A \vee B$'? Why don't we simply weigh the utility of A and B by their unconditional probability? Because then highly unlikely propositions would automatically have a utility near 0. If you are almost certain that the picnic will take place in Cedar Park, both Cr(*Alder Park*) and Cr(*Buckeye Park*) will be close to 0. But the mere fact that a proposition is unlikely does not make it undesirable. To evaluate the desirability of a proposition, we should bracket its probability. That's why Jeffrey's Axiom defines $U(A \vee B)$ the probability-weighted average of $U(A)$ and $U(B)$ *on the supposition that* $A \vee B$ *is true*.

Exercise 5.4 ⋆

You would like to win the lottery because that would allow you to travel the world, which you always wanted to do. Let *Win* be the proposition that you win the lottery, and *Travel* the proposition that you travel the world. Note that *Win* is logically equivalent to $(\mathit{Win} \wedge \mathit{Travel}) \vee (\mathit{Win} \wedge \neg \mathit{Travel})$, and thus has the same utility. Suppose $U(\mathit{Win} \wedge \mathit{Travel}) = 10$, $U(\mathit{Win} \wedge \neg \mathit{Travel}) = 0$, and your credence that you will travel the world on the supposition that you will win the lottery is 0.9. By Jeffrey's axiom, what is $U(\mathit{Win})$?

Exercise 5.5 ⋆⋆

At the beginning of this section, I argued that if $U(\neg A)$ is high, then $U(A)$ should be low, and vice versa. Let's use the utility of the tautology $A \vee \neg A$ as a neutral point of reference, so that $U(A \vee \neg A) = 0$. From this assumption, and Jeffrey's axiom, it follows that $U(\neg A) > 0$ just in case $U(A) < 0$. More precisely, it follows that

$$U(A)\mathrm{Cr}(A) = -U(\neg A)\mathrm{Cr}(\neg A).$$

Can you show how this follows? (It's not as hard as it looks. Hint: A is logically

incompatible with $\neg A$.)

Exercise 5.6 ⋆⋆

Let's say that an agent *desires* a proposition A just in case $U(A) > U(\neg A)$. One might have thought that whenever an agent desires a conjunction $A \wedge B$, then she should also desire A. But on the present understanding of desire, this is false. For example, if A is the proposition that I will break my leg in an accident today, and B is the proposition that I will get a billion pounds compensation, then I desire $A \wedge B$, but I do not desire A. What about the following hypotheses?

(a) Whenever an agent desires A, then she should also desire $A \vee B$.

(b) Whenever an agent desires A and desires B, then she should also desire $A \wedge B$.

Explain briefly and informally whether these are valid or invalid, perhaps by giving a counterexample like I did above.

Exercise 5.7 ⋆⋆⋆

Derive the following rule from Jeffrey's axiom and the rules of probability: if A, B, and C are incompatible with one another, then

$$U(A \vee B \vee C) = U(A) \cdot \mathrm{Cr}(A/(A \vee B \vee C)) + U(B) \cdot \mathrm{Cr}(B/(A \vee B \vee C)) + U(C) \cdot \mathrm{Cr}(C/(A \vee B \vee C)).$$

5.4 Basic desire

I have presented Jeffrey's axiom as the sole formal requirement on rational utility. Even that much is controversial. Many economists and philosophers hold that rationality imposes no constraints at all on an agent's desires. (In a way, this is the opposite extreme of the hedonist doctrine that all rational agents desire nothing but their own pleasure.) This idea was memorably expressed by David Hume in his *Treatise of Human Nature*:

> 'tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person unknown to me.

Hume claimed that basic goals are not responsive to evidence, reason, or argument. If your ultimate goal is to help some distant stranger, there is no non-circular argument that could prove your goal to be wrong, nor could we fault you for not taking into account any relevant evidence. Whatever facts you might find out about the world, you could coherently retain your ultimate goal of helping the stranger.

For Hume, beliefs and goals are in principle independent. What you believe is one thing, what you desire is another. Beliefs try to answer the question: what is the world like? Desires answer an entirely different question: what do you want the world to be like? On the face of it, these two questions really appear to be logically independent. Two agents could in principle give the same answer to the first question and different answers to second, or the other way around.

What we have seen in the previous section seems to contradict these intuitions. We have seen that an agent's utilities are thoroughly entangled with her credences. Indeed, we can read off an agent's credence in any proposition A from her utilities, assuming the utilities obey Jeffrey's axiom, the credences obey the probability axioms, and the agent is not indifferent towards A. For by Jeffrey's axiom,

$$U(A \vee \neg A) = U(A) \cdot \mathrm{Cr}(A) + U(\neg A) \cdot \mathrm{Cr}(\neg A).$$

By the Negation Rule, we can replace $\mathrm{Cr}(\neg A)$ by $1 - \mathrm{Cr}(A)$. Multiplying out, we get

$$U(A \vee \neg A) = U(A) \cdot \mathrm{Cr}(A) + U(\neg A) - U(\neg A)\mathrm{Cr}(A).$$

And then we can solve for $\mathrm{Cr}(A)$:

$$\mathrm{Cr}(A) = \frac{U(A \vee \neg A) - U(\neg A)}{U(A) - U(\neg A)}.$$

The ratio on the right-hand side is defined whenever $U(A) \neq U(\neg A)$.

What is going on here? Have we refuted Hume? Have we shown that an agent's beliefs are *part of her desires*?

Of course not – or not in any interesting sense. We need to distinguish **basic desires** from **derived desires**. If you are looking for a hammer to hang up a

picture, your desire to find the hammer is not a basic desire. Rather, it is derived from your desire to hang up the picture and your belief that you need a hammer to achieve that goal. By contrast, a desire to be free from pain is typically basic: if you want a headache to go away, this is usually not (or not only) because you think having no headache is associated with other things you desire. You simply don't want to have a headache, and that's the end of the story.

When Hume claimed that desires are independent of beliefs, he was talking about basic desires.

How are basic desires related to an agent's utility function? Imagine an agent whose *only* basic desire is to be free from pain, and let's pretend this is an all-or-nothing matter. The utility this agent gives to being free from pain then does not depend on her beliefs. All worlds in which she is free from pain are equally good, equally desirable, so it does not matter how the agent's credences are distributed among these worlds. Being pain-free *and rich*, for example, is equally desirable as being pain-free *and poor*. Both have the same utility as being pain-free itself.

Let's say that a proposition has *uniform utility* if the agent does not care how the proposition is realized: all (non-empty) subsets of the proposition (understood as a set of possible worlds) have equal utility.

What if an agent has two basic desires, say, being pain-free and being liked? These are logically independent, so there are four combinations: (1) being pain-free and liked, (2) being pain-free and not liked, (3) being in pain and liked, and (4) being in pain and not liked. Being pain-free no longer has uniform utility, since the worlds where the agent is pain-free divide into (better) worlds where the agent is pain-free and liked and (worse) worlds where the agent is pain-free and disliked. As a consequence, the utility of being pain-free now depends on the agent's beliefs: the stronger she believes that she is liked if she is pain-free, the more she desires being pain-free.

The four combinations of being pain-free and being liked, however, have uniform utility. All worlds in which the agent is, say, pain-free and not liked are equally desirable (pretending these are all-or-nothing features). Let's call such a combination a **concern**. So a concern is a proposition that specifies everything the agent ultimately cares about.

We can think of an agent's utility function as built up from the utility she attaches to her concerns. The choice of concerns, and the utilities they get, is independent of the agent's beliefs. Once the utilities of the agent's concerns are fixed, the agent's credence function determines the utility of all other propositions, by Jeffrey's axiom.

Exercise 5.8 ⋆⋆

There's a party, and at first you want to be invited. Then you hear that Bob will be there, and you no longer want to be invited. Then you hear that there will be free beer, and you want to be invited again. Your desire seems to change back and forth. Nonetheless, your basic desires may have remained the same throughout. Explain how your fluctuating attitude might have come about without any change in basic desires.

5.5 Further reading

For an eloquent defence of (roughly) our approach to utility, with tangents into social and political issues, read chapter 6 ("Game Theory and Rational Choice") of

- Simon Blackburn: *Ruling Passions* (1998).

Along similar lines, but shorter and a little more technical:

- John Broome: "'Utility'" (1991).

The formal theory of utility reviewed in section 5.3 comes from chapter 5 of

- Richard Jeffrey: *The Logic of Decision* (1965/1983).

Essay Question 5.1

Do you agree with Hume that there are no rational constraints on basic desires? If so, try to defend this view. If not, try to argue against it.

6 Preference

6.1 The ordinalist challenge

If the utility of an outcome for an agent is not measured by the amount of money the agent gains or loses, how is it measured? How can we find out whether an outcome has utility 5 or 500 or -27? What does it even mean to say that an outcome has utility 5?

At the beginning of the 20th century, doubts arose about the coherence of numerical utilities. **Ordinalists** like Vilfredo Pareto argued that the only secure foundation for utility judgements are people's choices: if you are given a choice between tea and coffee, and you choose tea, we can conclude that tea has greater utility for you than coffee. We may similarly find that you prefer coffee to milk, etc., but how could we find that your utility for tea is twice your utility for coffee – let alone that it has the exact value 5? The ordinalists concluded that we should give up the conception of utility as a numerical magnitude.

Ordinalism posed a serious threat to the idea of expected utility maximization. If there is no numerical quantity of utility, we can't demand that rational agents maximize the probability-weighted average of that quantity, as the MEU Principle requires.

In 1926, Frank Ramsey pointed out that if we look at the choices an agent makes in a state of uncertainty, we may discover more about the agent's utility function than how it orders the relevant outcomes – enough to vindicate the MEU Principle. Ramsey's line of thought was largely ignored until it was rediscovered by John von Neumann, who presented a simpler version of it in the 1944 monograph *Game Theory and Economic Behaviour*, co-authored with Oskar Morgenstern. This work is widely taken to provide the foundation of modern expected utility theory.

Before we have a closer look at von Neumann's argument, let's think a little about the ordinalist challenge.

Ordinalism was inspired by a wider "positivist" movement in science and philosophy whose aim was to improve scientific reasoning by discarding reference

to seemingly obscure and unobservable facts. According to positivism, any meaningful statement must have clear conditions of verification and falsification. If someone puts forward a hypothesis but can't explain how one could in principle test whether the hypothesis is true or false, then the hypothesis should be rejected as meaningless. In psychology, this movement gave rise to **behaviourism**, the doctrine that statements about emotions, intentions, desires, and other psychological states should either be abandoned or defined in terms of observable behaviour.

Today, behaviourism, and positivism more generally, have been almost entirely abandoned. One reason for this (to which I already alluded in the previous chapter) is that people came to appreciate the holistic character of scientific testing: many statements in successful scientific theories have observable consequences only in conjunction with other theoretical assumptions. More practically, the behaviourist paradigm was found to stand in the way of scientific progress. It is hard to explain even the behaviour of simple animals without appealing to inner representational states like goals or perceptions as causes of the behaviour.

On the basis of these historical developments, it may be tempting to dismiss the ordinalist challenge as misguided. However, even if their general view of science was mistaken, the ordinalists raised an important issue.

In chapter 3, I emphasized that we should not think of an agent's credences as little numbers written in her head. If your credence in rain is $1/2$, then this must be grounded in other, more basic facts about you – facts that do not involve the number $1/2$. Even if we accept your state of belief as a genuine internal state, a cause of your behaviour, we need to explain why we represent that state by the number $1/2$ rather than $3/4$ or $12/5$. There's nothing special here about credence. Numerical representations in scientific models are always based on non-numerical facts about the represented objects. For the numerical representations to have meaning, we need to specify what underlying non-numerical facts the different numbers are meant to represent.

The same is true for utility. As I explained in the previous chapter, we understand utility to comprise a wide range of psychological factors, from unconscious aversions to cravings to moral judgements. What unites all these factors is that they make the relevant proposition more or less appealing to the agent. The utility of a proposition is supposed to represents the extent to which the agent, on balance, wants the proposition to be true – taking into account all the agent's attitudes towards the proposition. But it really isn't obvious why a given combination of pro- and contra-attitudes should be represented by the number 5, say,

as opposed to any other number.

So the ordinalists raised a question that still needs to be answered. Moreover, there is something to be said for the idea that the answer should involve the agent's choices.

Consider our practice of attributing conscious or unconscious motives. A child bullies other children at school. Why does she do this? One hypothesis is that she simply enjoys the sense of power. Another is that her aggression is an attempt to hide her insecurities and protect herself from an unconsciously perceived threat posed by other children. A minimal standard for any such explanation is that the goals it postulates make sense of the child's behaviour. That is, on the assumption that the child is motivated by the hypothesized factors, we would expect to see the kind of behaviour we actually observe.

In general, the main reason to think that an agent has specific goals or desires is that this would explain her behaviour. The point also applies to the relative strength of the attributed goals or desires. We say that my sense of duty was stronger than my desire to stay in bed because I actually got up. Absent further explanation, the claim that my desire to stay in bed was stronger, even though I got up, is unintelligible.

So there is a close connection between an agent's motives or goals or desires, and her behaviour. What it means to be in a particular motivational state is, at least in part, to be in a state that typically leads to particular kinds of behaviour, given suitable beliefs. If we seek a standard to measure the comparative strength of different motives, a natural move is to look at their behavioural consequences.

6.2 Scales

Utility, like credence, mass, or length, is a numerical representation of an essentially non-numerical phenomenon. All such representations are to some extent conventional. We can represent the length of my pencil as 19 centimetres or as 7.48 inches – it's the same length either way. We must take care to distinguish real features of the represented properties from arbitrary consequences of a particular representation. For example, it is nonsense to ask whether the length of my pencil – the length itself, not the length in any particular system of representation – is a whole number. By contrast, it is not meaningless to ask whether the length of my pencil is greater than the length of my hand.

The mathematical discipline of measure theory studies the representation of

physical properties by numbers. In the case of length, as in the case of mass, the conventionality boils down to the choice of a unit. You can introduce a new measure of length simply by picking out a particular length and assigning it the number 1. Any object twice that length will then have a length of 2 in your new system, and so on. (You can fix the unit by assigning any number greater than zero to any non-trivial length; it doesn't have to be the number 1.)

Quantities like mass and length, for which only the unit of measurement is conventional, are said to have a **ratio scale**, because even though the particular numbers are conventional, ratios between them are not: if the length of my arm is four times the length of my pencil in centimetres, then that is also true in inches, millimetres, and any other sensible system of measurement.

Temperature is different. Historically, the basis for representing temperature by numbers was the observation that metals such as mercury expand as the temperature goes up. Imagine we put some mercury in a narrow glass tube. The higher the temperature, the more of the glass tube is filled up by the expanding mercury. To get a numerical measure of temperature, we need to mark two points on the tube – for example, by 0 and 100. We can then say that if the mercury has expanded to x% of the distance between 0 and 100, then the temperature is x. Anders Celsius suggested to use 0 for the temperature at which water freezes, and 100 for the temperature at which it boils. Daniel Fahrenheit instead marked as 0 the coldest temperature he measured in his home town of Danzig in the winter of 1708/1709, and used 100 for the body temperature in a healthy human. As a result, 10 degrees Celsius is 50 degrees Fahrenheit, and 20 degrees Celsius is 68 degrees Fahrenheit. Since the ratio between the two temperatures is not preserved, the Celsius scale and the Fahrenheit scale are not ratio scales. Such scales, where both the zero and the unit are a matter of convention, are called **interval scales**.

The ordinalists held that utility has neither a ratio scale nor an interval scale, but merely an **ordinal scale** (hence the name of the movement). In an ordinal scale, the only thing that is not conventional is which of two objects is assigned a greater number. For example, according to the ordinalists, a preference of tea over coffee and of coffee over milk can be represented by a utility function that assigns 3 to tea, 2 to coffee, and 1 to milk, but it can also be represented by a utility function that assigns 300 to tea, 0 to coffee, and -1 to milk. Either function correctly reflects your choices.

If the ordinalists were right, we would have to give up the MEU Principle. Which act in a decision problem maximizes expected utility would then frequently

depend on arbitrary conventions for representing utility.

By contrast, if utility has an interval scale, then different measures of utility will never disagree on the ranking of acts in a decision problem. A ratio scale is not required.

Exercise 6.1 ⋆

In the Mushroom Problem as described by the matrix on page 12 (section 1.3), not eating the mushroom has greater expected utility than eating the mushroom. Describe a different assignment of utilities to the four outcomes which preserves their ordering but gives eating the mushroom greater expected utility than not eating.

Exercise 6.2 ⋆⋆

Suppose two utility functions U and U' differ merely by their choice of unit and zero. It follows that there are numbers $x > 0$ and y such that, for any A, $U(A) = x \cdot U'(A) + y$. Suppose some act A in some decision problem has greater expected utility than some act B, if the utility of the outcomes is measured by U. Show that A also has greater expected utility than B if the utility of the outcomes is measured by U'. (You can assume for simplicity that the outcome of either act depends only on whether some state S obtains; so the states are S and $\neg S$.)

So if we want to rescue the MEU Principle from ordinalist skepticism, we don't need to explain what makes it the case that your utility for tea is 5 rather than 500; we can accept that the precise numbers are a matter of conventional representation. Nor do we need to explain what makes your utility for tea twice your utility for coffee; such ratios also needn't track anything real. But we do have to explain what makes it the case that once we arbitrarily set your utility for tea as 5 and your utility for coffee as 0, then your utility for milk is fixed at, say, -7.

6.3 Utility from preference

I am now going to describe John von Neumann's proposal for how to determine an agent's utility function from her choice dispositions.

Recall from the previous chapter that an agent's utility function is fixed by

the utility assigned to the propositions I called *concerns*. Intuitively, a concern settles everything that matters to the agent, leaving open only questions towards which the agent is indifferent. To make the following discussion a little more concrete (and to bypass some problems that will occupy us later), let's imagine an agent who is only interested in getting certain "rewards", which may be lumps of money or commodity bundles or pleasant sensations. I will use lower-case letters $a, b, c, \ldots$ for rewards. Our goal is to find the agent's numerical utility for $a, b, c, \ldots$.

We will determine the agent's utilities from her **preferences**, which we assume to represent her choice dispositions. For example, let's say the agent would choose reward a if she were given a choice between a and b. She then prefers a to b. The ordinalists did not challenge the idea that people have preferences in this sense.

Let's introduce some shorthand notation:

$a \succ b \iff$ The agent prefers a to b.

$a \sim b \iff$ The agent is indifferent between a and b.

$a \succsim b \iff$ The agent prefers a to b or is indifferent between them.

(Note that '$\succ$', '$\sim$', and '$\succsim$' had a different meaning in section 3.6. You always have to look at the context to figure out what these symbols mean.)

We saw that to defend the MEU Principle, we can choose an arbitrary unit and zero for the utility scale. So let's take arbitrary rewards a and b such that $b \succ a$ and set $U(a) = 0$ and $U(b) = 1$. This corresponds to Kelvin choosing zero as the temperature at which water freezes and 100 as the temperature at which water boils.

Exercise 6.3 ⋆⋆

If the agent is indifferent between all rewards, then our procedure stalls at this step. Nonetheless, we can easily find a utility function for such an agent. What does it look like?

Having fixed the utility of two rewards a and b, here is how we can determine the utility of any other reward c. We distinguish three cases, depending on how the agent ranks c relative to a and b.

Suppose first that c "lies between" a and b in the sense that $b \succ c$ and $c \succ a$. To find the utility of c, we look at the agent's preferences between c and a **lottery** between a and b. By a 'lottery between a and b', I mean an event that leads to a

with some objective probability x and otherwise to b. For example, suppose we offer our agent a choice between c for sure and the following gamble L: we'll toss a fair coin; on heads the agent gets a, on tails b. By the Probability Coordination Principle, the expected utility of the gamble is $1/2$. So if the agent is indifferent between the lottery and c, and the agent obeys the MEU Principle, we can infer that the agent's utility for c is also $1/2$.

Exercise 6.4 ⋆⋆

Suppose an agent's utility is 0 for a, 1 for b, and $1/2$ for c. Draw a decision matrix representing a choice between c and L, and verify that both options have expected utility $1/2$.

Exercise 6.5 ⋆⋆

Why do we need to assume that the agent obeys the Probability Coordination Principle?

If the agent isn't indifferent between L and c, we try other lotteries, until we find one for which the agent is indifferent between the lottery and c. For example, suppose the agent is indifferent between c and a gamble L' where she would get a with probability $4/5$ and b with probability $1/5$. Since the expected utility of this gamble is $1/5$, we could then infer that the agent's utility for c is $1/5$.

We have assumed that c lies between a and b. What if the agent prefers c to both a and b? Then we will look for a lottery between a and c such that the agent is indifferent between b and the lottery. For example, if the agent is indifferent between b for sure and a gamble L'' where she gets either a or c with equal probability, then c must have utility 2. That's because the expected utility of L'' is

$$EU(L'') = 1/2 \cdot U(a) + 1/2 \cdot U(c) = 0 + 1/2 \cdot U(c) = 1/2 \cdot U(c).$$

Since the agent is indifferent between L'' and b, which has a guaranteed utility of 1, the gamble must have expected utility 1. So $1 = 1/2 \cdot U(c)$. And so $U(c) = 2$. In general, if the agent is indifferent between b and a lottery that leads to c with probability x and a with probability $1 - x$, then $U(c) = 1/x$.

Exercise 6.6 ⋆⋆

Can you complete the argument for the case where the agent prefers both a and b to c?

In this manner, we can determine the agent's utility for all rewards from her preferences between rewards and lotteries. And so we can answer the ordinalist challenge: we can define the agent's utility as whatever utility function we could read off in this way from her preferences. This is the official definition of 'utility' in most economics textbooks. (A simpler definition of a merely ordinal concept of utility in terms of preferences is also frequently used for applications where uncertainty can be ignored.)

6.4 The von Neumann and Morgenstern axioms

The method described in the previous section assumes that the agent obeys the MEU Principle. At first glance, that may seem strange. The ordinalists argued that the MEU Principle made no sense; how can we respond to them by *assuming* the principle? Besides, doesn't application of the MEU Principle presuppose that we already know the agent's utilities?

The trick is that we are applying the principle backwards. Normally, when we apply the MEU Principle, we start with an agents beliefs and desires and try to find out her (optimal) choices. Now we start with her choices and try to find out her desires, relying on the Probability Coordination Principle to fix the relevant beliefs.

There is nothing dodgy about this. Whenever we want to measure a quantity whose value can't be directly observed, we have to rely on assumptions about how the quantity relates to other things that we can observe. Together with the Probability Coordination Probability, the MEU Principle tells us what lotteries an agent should be disposed to accept if she has a given utility function. If she wouldn't accept those lotteries, we can infer that she doesn't have the utility function, and so we look at other lotteries until we find a match.

You may wonder, though, what happened to the normativity of the MEU Principle. If we follow von Neumann's method to define an agent's utility function, won't the agent automatically come out as obeying the MEU Principle, at least for the relevant lotteries and rewards?

Not quite. It's true that *if the method* works, then the agent will evaluate *certain* lotteries by their expected utility, relative to the utility function identified by the method. But the method is not guaranteed to work, and it does not settle how the agent evaluates arbitrary lotteries.

To illustrate the first point, we have assumed that if an agent ranks some

reward c in between a and b, then the agent is indifferent between c and some lottery between a and b. That is not a logical truth. An agent could in principle prefer c to any lottery between a and b, yet still prefer c to a and b to c. Von Neumann's method does not identify a utility function for such an agent.

Von Neumann and Morgenstern investigated just what conditions an agent's preferences must satisfy in order for the method to work, and for it to guarantee that the agent evaluates arbitrary lotteries by their expected utility. To state these conditions, I will assume that '$\succ$', '$\sim$', and '$\succsim$' are defined not just for basic rewards but also for lotteries between rewards as well as "compound lotteries" whose payoff is another lottery. For example, if I toss a fair coin and offer you lottery L on heads and L' on tails, that would be a compound lottery.

Here are the conditions we need. 'A', 'B', 'C' range over arbitrary lotteries or rewards.

Completeness

For any A and B, exactly one of $A \succ B$, $B \succ A$, or $A \sim B$ is the case.

Transitivity

For any A, B, and C, if $A \succsim B$ and $B \succsim C$ then $A \succsim C$.

Continuity

For any A, B, and C, if $A \succ B$ and $B \succ C$ then there are lotteries L_1 and L_2 between A and C such that $L_1 \succ B$ and $B \succ L_2$.

Independence (of Irrelevant Alternatives)

For any A, B, and C, if $A \succ B$, and L_1 is a lottery that leads to A with some probability x and otherwise to C, and L_2 is a lottery that leads to B with probability x and otherwise to C, then $L_1 \succ L_2$.

Reduction (of Compound Lotteries)

If a L_1 and L_2 are two (possibly compound) lotteries that lead to the same rewards with the same objective probabilities, then $L_1 \sim L_2$.

Von Neumann and Morgenstern proved that if (and only if) an agent's preferences satisfy all these conditions, then there is a utility function U, determined by the method from the previous section, which *represents* the agent's prefer-

ences in the sense that $A \succ B$ just in case $U(A) > U(B)$, and $A \sim B$ just in case $U(A) = U(B)$. Moreover, the function U is unique except for the choice of unit and zero. (That is, any two functions U and U' that represent the agents preferences differ at most in their choice of unit and zero.) This result is known as the **von Neumann-Morgenstern Representation Theorem**.

So if we follow von Neumann and Morgenstern's definition of utility, then the MEU Principle (for choices involving lotteries) will automatically be satisfied by any agent whose preferences satisfy the above conditions – Completeness, Transitivity, etc. The normative claim that an agent ought to evaluate lotteries by their expected utility therefore reduces to the claim that their preferences ought to satisfy the conditions. For this reason, the conditions are also known as the **axioms of expected utility theory**.

Von Neumann and Morgenstern therefore offered not only a response to the ordinalist challenge. They also offered a powerful argument for the MEU Principle. The argument could be spelled out as follows.

1. The preferences of a rational agent satisfy Completeness, Transitivity, Continuity, Independence, and Reduction.
2. If an agent's preferences satisfy these conditions, then (by the Representation Theorem) they are represented by a utility function U relative to which the agent ranks options by their expected utility.
3. That utility function U is the agent's true utility function.
4. Therefore: a rational agent ranks options by their expected utility.

Exercise 6.7 ⋆

(An example due to John Broome.) Maurice would choose to go to Rome if he were offered a choice between Rome and going to the mountains, because the mountains frighten him. Offered a choice between staying at home and going to Rome, he would prefer to stay at home, because he finds sightseeing boring. But if he were offered a choice between mountaineering and staying at home, he would choose the mountains because it would be cowardly, he believes, to stay at home. Which of the axioms does Maurice appear to violate?

6.5 Utility and credence from preference

In chapter 3 we looked at the betting interpretation, which attempts to derive an agent's credences from her choice dispositions. We saw that the approach relied on implausible assumptions about the agent's utility function. Meanwhile, we have learned from von Neumann and Morgenstern how we might derive an agent's utility function from her choice dispositions. Ramsey (in 1926) argued that the two tasks can be combined. He showed how we might simultaneously determine both an agent's credences and her utilities from her choices. Ramsey's idea was rediscovered and streamlined by Leonard Savage in his *Foundations of Statistics* (1954) – the second most influential book in the history of decision theory, after *Game Theory and Economic Behaviour*. I will briefly describe Savage's main result.

Like von Neumann and Morgenstern, Savage begins with some conditions ("axioms") on an agent's comparative preference relation, which we assume to reflect the agent's choice dispositions. This time, the preference relation is defined over a set of basic rewards as well as *conditional prospects*. A conditional prospect is an event that leads to some reward a if some state of the world X obtains and otherwise to a possibly different reward b. I will abbreviate such a prospect by '$[X\,?\,a : b]$' (pronounced 'if X then a else b'). We also allow for conditional prospects in which the outcomes are not basic rewards but further conditional prospects. The intuitive thought is that any act in any decision problem corresponds to a conditional prospect. In the mushroom problem from chapter 1, for example, eating the mushroom amounts to choosing the prospect [Poisonous ? Dead : Satisfied]; not eating the mushroom amounts to choosing [Poisonous ? Hungry : Hungry].

Savage's axioms are a little more complicated than those of von Neumann and Morgenstern. As before, we need Completeness and Transitivity – I won't repeat them. We also need to assume that the agent is not indifferent between all rewards, as then we would have no means of discovering her beliefs:

Non-Triviality

There are rewards a and b for which $a \succ b$.

The Independence axiom is redefined as follows, to reflect the change from lotteries to conditional prospects.

Independence (of Irrelevant Alternatives)

If two prospects A and B lead to different outcomes only under condition X, then for any C, $A \succ B$ iff $[X\,?\,A : C] \succ [X\,?\,B : C]$.

Instead of Continuity and Reduction we have the following conditions. (Don't worry if you can't make immediate sense of them.)

Nullity

If $A \succ B$ and $[X\,?\,A : C] \not\succ [X\,?\,B : C]$, then for all A', B', $[X\,?\,A' : C] \not\succ [X\,?\,B' : C]$.

Stochastic Dominance

If $A \succ B$ and $A' \succ B'$, then $[X\,?\,A : B] \succ [Y\,?\,A : B]$ iff $[X\,?\,A' : B'] \succ [Y\,?\,A' : B']$.

State Richness

If $A \succ B$, then for all C there is a finite number of mutually exclusive and jointly exhaustive states $X_1, \ldots, X_n$ such that for all X_i, $[X_i\,?\,C : A] \succ B$ and $A \succ [X_i\,?\,C : B]$.

Averaging

It is not the case that for all outcomes O that might result from prospect A under condition X, $[X\,?\,A : B] \succ [X\,?\,O : B]$, nor is it the case that $[X\,?\,O : B] \succ [X\,?\,A : B]$ for all such outcomes.

Savage's Representation Theorem states that if (but not only if) an agent's preferences satisfy all these conditions, then there is a utility function U and a probability function Cr such that $A \succ B$ iff the expected utility of A, relative to Cr and U, is greater than that of B. Moreover, Cr is unique and U is unique expect for the choice of zero and unit.

What can this do for us? Well, suppose we *define* an agent's credence and utility functions as the functions Cr and U whose existence and uniqueness (modulo conventional choice of zero and unit, for U) is guaranteed by Savage's theorem, provided the agent's preferences satisfy the axioms. If we accept the axioms as genuine norms of rationality, we can then explain what non-numerical facts the credences and utilities of rational agents are meant to represent: they represent certain patterns in the agent's preferences and therefore ultimately in

her choice dispositions. For on the present approach, saying that a rational agent has credences Cr and utilities U is equivalent (by the proposed definition of Cr and U) to a certain claim about the agent's preferences. We no longer need the betting interpretation, and we have answered the ordinalist challenge.

In addition, the present approach promises a more comprehensive argument for the MEU Principle than the argument we got from von Neumann and Morgenstern. Their argument only showed that agents should rank *lotteries* by their expected utility. But not all choices involve lotteries. In real life, agents often face conditional prospects in which they don't know the objective probability of the various outcomes. Why should they rank such prospects by their expected utility? Savage's answer is that if they don't, then they don't satisfy his axioms.

We also get a new argument for probabilism – the claim that rational degrees of belief satisfy the probability axioms. Again, the requirement reduces to the preference axioms: on the proposed definition of credence, any agent who obeys these axioms automatically has probabilistic credences. If you don't have probabilistic credences, you violate the axioms.

Exercise 6.8 ⋆⋆

Can you spell out the argument for probabilism I just outlined in more detail, in parallel to the argument for the MEU Principle I spelled out at the end of section 6.4?

It should now be clear why the results of Savage and von Neumann and Morgenstern are widely taken to provide the foundations of expected utility theory. The results not only seem to show how an agent's credence and utility function can be measured in terms of overt choices, they also suggest that our main norms – probabilism and the MEU Principle – reduce to certain conditions on choices. To finish the job, it seems, we only have to convince ourselves that these conditions (the "axioms") are genuine norms of rationality.

In fact, doubts have been raised about every single one of the axioms. We will turn to some of these worries in chapter 8. In the meantime, I want to flag a different kind of problem.

6.6 Preference from choice?

Von Neumann and Savage take as their starting point an agent's preferences, represented by the relations $\succ$, $\sim$, and $\succsim$. Informally, we have interpreted '$A \succ B$'

as stating that the agent would choose A if she were offered a choice between A and B. The representation theorems are then supposed to explain how an agent's utilities (or utilities and credences) can be measured by her choice dispositions.

An agent's *dispositions* reflect what she *would* do if such-and-such circumstances were to arise. It should be clear that we can't just look at the agent's actual choice behaviour, since most agents are not confronted with all the choices from which von Neumann or Savage would derive their utility function.

Exercise 6.9 ⋆⋆

Suppose we define '$A \succ B$' as 'the agent has been confronted with a choice between A and B, and chose A'; similarly for '$A \sim B$' and '$A \succsim B$'. Which of the von Neumann and Morgenstern axioms then become highly implausible?

But now one of the problems for the betting interpretation, from section 3.4, returns with a vengeance. If an agent is in fact not facing a choice between two options A and B, then offering her the choice would change her beliefs. Among other things, she would come to believe that she faces that choice. According to the MEU Principle, the agent in the hypothetical choice situation should choose whichever option maximizes expected utility relative to the beliefs and desires she has in that situation. So if we interpret preferences in terms of hypothetical choices, then we cannot assume that a rational agent prefers A to B just in case A has greater expected utility than B relative to the agent's actual beliefs and desires.

The problem gets worse if we drop the simplifying assumptions that agents only care about lumps of money, commodity bundles, or pleasant sensations. Suppose one thing you desire (one "reward") is peace in Syria, another is being able to play the piano. Von Neumann's definition then determines your utilities in part by your preferences between peace in Syria and a lottery that leads to peace in Syria with objective probability $1/4$ and to an ability to play the piano with probability $3/4$. Savage's method will similarly look at your preferences between peace in Syria and various prospects like $[\mathit{Rain}\,?\,\mathit{Peace} : \mathit{Piano}]$ – an imaginary act that leads to peace in Syria if it rains and to an ability to play the piano if it doesn't rain. If you thought you'd face this bizarre choice, your beliefs would surely be quite different from your actual beliefs. Indeed, no matter what you pick in the hypothetical choice situation, it is guaranteed that there is either peace in Syria or you can play the piano. So you could only believe that you face the hypothetical choice if you are sure one of these propositions is true. Clearly we

can't assume that your credences in the hypothetical choice situation match your actual credences.

Even in the rare case where an agent actually faces one of the relevant choices, we arguably can't infer that whichever option she chooses has greater expected utility for her.

For one thing, people can have false beliefs about their options. If your real choice is between water and wine, but you think it is between petrol and wine (because you think the water is petrol), we can't infer from your choice of wine that you prefer wine to water. Von Neumann and Savage presuppose that agents are never mistaken about their options.

Moreover, if an agent chose A over an alternative B, we can't infer that she genuinely preferred A. Perhaps she was indifferent and chose A merely because she had to make a choice. Arguably, choice dispositions therefore can't tell apart $A \succ B$ and $A \sim B$.

The upshot of all these problems is that we need to distinguish (at least) two notions of preference. One represents the agent's choice dispositions: whether she would choose A over B in a hypothetical situation in which she faces that choice. The other represents the agent's current ranking of hypothetical prospects or lotteries: whether by the lights of her current beliefs and desires, A is better than B. Von Neumann and Savage at best demonstrated how to derive utilities and credences from preferences in the second sense.

This could still be valuable. For example, we might still get interesting arguments for probabilism and the MEU Principle. Moreover, there is plausibly *some* connection between preference in the second sense and choices dispositions, so even though we haven't fully solved the measurement problem for credences and utilities, one might hope that we are at least a few steps closer.

6.7 Further reading

For some recent discussion of the idea that utility and credence are derived from preferences, see (for example)

- Samir Okasha: "On the Interpretation of Decision Theory" (2016),
- Daniel Hausman: "Mistakes about Preferences in the Social Sciences" (2011).

You may notice that Okasha and Hausman mean different things by 'preference'.

A useful survey of many further representation theorems in the style of those we have reviewed is

- Peter Fishburn: "Utility and Subjective Probability" (1994).

Essay Question 6.1

Do you think an agent's choice dispositions can in principle reveal all her goals and values? If yes, can you explain how? If no, can you explain why not?

7 Separability

7.1 The construction of value

At the end of chapter 5 we saw that the utility of a proposition for an agent is determined by two factors: the agent's credences, and the agent's basic desires, which are reflected in the utility assigned to the agent's "concerns".

For example, suppose all you ultimately care about is that you and your friends are happy, and let's pretend this is an all-or-nothing matter. All worlds in which you and your friends are happy then are equally desirable for you, no matter what else happens at these worlds. In the terminology of chapter 5, the set of these worlds has "uniform utility". The same is true for the worlds in which you are happy and your friends are unhappy, and for the worlds in which you are unhappy and your friends are happy, and for the worlds in which you and your friends are all unhappy. These four sets of worlds are your "concerns".

In general, if there are n propositions $A_1, A_2, \ldots, A_n$ that you ultimately care about, then any conjunction that can be formed from these propositions and their negations (such as $A_1 \wedge \neg A_2 \wedge A_3 \wedge \ldots \wedge \neg A_n$) has uniform utility, and is one of your concerns.

If we know an agent's credence function, and the utility she assigns to her concerns, we can use Jeffrey's axiom to compute the agent's utility for any proposition. The utilities assigned to the agent's concerns represent the agent's basic, belief-independent desires, whereas the utilities assigned to other propositions typically represent a combination of basic desires and beliefs.

In this chapter, we will look at how the utility of a concern might be determined. It will be useful to have a label for an agent's utility function restricted to concerns. I will call it the agent's **value function**. So an agent's value function specifies to what extent the agent desires all the combinations of propositions she ultimately cares about. The value function represents the agent's belief-independent desires.

7.2 Additivity

Let's start with a toy example. You are looking for a flat to rent. You care about various aspects of a flat such as size, location, and price. We'll call these aspects **attributes**. If a set of attributes comprises all the features (of a flat) that matter to you, then your preferences between possible flats are determined by your preferences between combinations of these attributes: if you prefer one flat to another, that's because you prefer the combined attributes of the first to those of the second.

So the desirability of any possible flat is determined by the desirability of every possible combination of attributes. We'll write these combinations as lists enclosed in angular brackets. For example, '$\langle 40\text{m}^2, \text{central}, £500\rangle$' stands for any flat with a size of 40 m^2, central location, and monthly costs of £500. If these are all the attributes you care about, then your utility function will assign the same value to all such flats.

It's the same with possible worlds. If all you care about is the degree of pleasure of you and your three best friends, then we can represent your basic desires by a value function that assigns numbers to lists like $\langle 10, 1, 2, 3\rangle$, specifying degrees of pleasure for you and your friends (in some fixed order). Each such list effectively specifies one of your concerns: a maximal conjunction of propositions you care about.

Return to the flats. Assuming you only care about size, location, and price, there will be some value function that assigns a desirability score to possible combinations of size, location, and price. If you're like most people, we can we say more about how these scores are determined on the basis of the individual attributes. For example, you probably prefer cheaper flats to more expensive flats, and larger ones to smaller ones.

A natural method for determining the overall score for a given flat goes as follows. First, assign scores to each attribute of the flat. Then add up these scores. For example, a cheap but small flat in a good location gets a high score for price, a low score for size, and a high score for location, which amounts to a medium overall score.

More formally, the present idea is to define your value for any given attribute list as the sum of **subvalues** assigned to individual attributes in the list. That is, if $V_S(40\text{m}^2)$ is the score you assign to a size of 40 m^2, $V_L(\text{central})$ is the score for

central location, and $V_P(£500)$ is the score for monthly costs of £500, then

$$V(\langle 40\text{m}^2, \text{central}, £500\rangle) = V_S(40\text{m}^2) + V_L(\text{central}) + V_P(£500).$$

If a value function V is determined by adding up subvalues in this manner, then V is called **additive** relative to the attributes in question.

Additivity may seem to imply that you assign the same weight to all the attributes: that size, location, and price are equally important to you. To allow for different weights, you might think, we should include scaling factors w_S, w_L, w_P, like so:

$$V(\langle 40\text{m}^2, \text{central}, £500\rangle) = w_s \cdot V_S(40\text{m}^2) + w_L \cdot V_L(\text{central}) + w_P \cdot V_P(£500).$$

However, we can get omit the weights by folding them into the subvalues. That is, we will let $V_S(200\text{m}^2)$ measure not just how awesome it would be to have a 200 m^2 flat, but also how important this feature is compared to price and location.

Exercise 7.1 ⋆

Like utility functions, subvalue functions assign numbers to sets of possible worlds that may vary in desirability. But unlike utility functions, subvalue functions are insensitive to belief. This explains why, if you can afford to pay £600 in monthly rent, then $V_P(£300)$ is plausibly high, even though the *utility* you assign to renting a flat for £300 is low. Can you spell out the explanation?

Exercise 7.2 ⋆⋆⋆

Additivity greatly simplifies an agent's psychology. Suppose an agent's basic desires pertain to 10 propositions $A_1, A_2, \ldots, A_{10}$. There are $2^{10} = 1024$ conjunctions of these propositions and their negations (such as $A_1 \wedge A_2 \wedge \neg A_3 \wedge \neg A_4 \wedge A_5 \wedge A_6 \wedge \neg A_7 \wedge A_8 \wedge A_9 \wedge \neg A_{10}$). To store the agent's value function in a database, we would therefore need to store up to 1024 numbers. By contrast, how many numbers would we need to store if the value function is additive?

7.3 Separability

Under what conditions is value determined by adding subvalues? How are different subvalue functions related to one another? What do subvalue functions

represent anyway? We can get some insight into these questions by following an idea from the previous chapter and study how an agent's value function might be derived from her preferences.

For the moment, we want to set aside the influence of the agent's beliefs, so we are not interested in an agent's preferences between lotteries or conditional prospects. Rather, we will look at an agent's preferences between complete attribute lists, assuming the relevant attributes comprise everything the agent cares about.

The main motivation for starting with preferences is, as always, the problem of measurement. We need to explain what it means that your subvalue for a given attribute is 5 or 29. Since the numbers are supposed to reflect, among other things, the importance (or weight) of the relevant attribute in comparison to other attributes, it makes sense to determine the subvalues from their effect on overall rankings.

So assume we have preference relations $\succ$, $\succsim$, $\sim$ between some lists of attributes. To continue the illustration in terms of flats, if you prefer a central 40 m^2 flat for £500 to a central 60 m^2 for £800, then

$$\langle 40\text{m}^2, \text{central}, £500\rangle \succ \langle 60\text{m}^2, \text{central}, £800\rangle.$$

Above we've assumed that you prefer cheaper flats to more expensive flats, so that V_P is a decreasing function of the monthly costs: the higher the costs c, the lower $V_p(c)$. But of course you don't prefer *any* cheaper flat to *any* more expensive flat. You probably don't prefer a 5 m^2 flat for £499 to a 60 m^2 flat for £500. The other attributes also matter.

In what sense, then, do you prefer cheaper flats to more expensive flats? We can cash this out as follows: whenever two flats agree in terms of size and location, and one is cheaper than the other, then you prefer the cheaper one.

Let's generalize this concept. Suppose A_1 and A_1' are two attributes that can occur in the first position of an attribute list – for example 40 m^2 and 60 m^2 if the first position represents the size of a flat, or £499 and £500 if it represents price, etc. For any way of filling in all other positions in the list, your preferences between attribute lists determine a ranking of A_1 and A_1'. Call these your preferences between A_1 and A_1' *conditional on* the attributes in the other positions. That is, if

$$\langle A_1, A_2, \ldots, A_n\rangle > \langle A_1', A_2, \ldots, A_n\rangle,$$

then you prefer A_1 to A_1' conditional on $A_2, \ldots, A_n$. Now suppose your preferences between A_1 and A_1' are the same conditional on any way of filling in the other

attributes. That is, if $\langle A_1, A_2, \ldots, A_n \rangle > \langle A'_1, A_2, \ldots, A_n \rangle$, and we replace $A_2, \ldots, A_n$ by arbitrary alternatives $A'_2, \ldots, A'_n$, we still get $\langle A_1, A'_2, \ldots, A'_n \rangle > \langle A'_1, A'_2, \ldots, A'_n \rangle$. In that case, let's say that your preferences between A_1 and A'_1 are *independent* of the other attributes.

In the example of the flats, your preference of £400 over £500 is plausibly independent of the other attributes, for whenever two possible flats agree in size and location, but one costs £400 and the other £500, you prefer the one for £400.

Now suppose your preferences between any two attributes in the first position (not just A_1 and A'_1) are independent of the other attributes. Moreover, suppose your preferences between any attributes in the second position are independent of the other attributes. And so on for all positions. Then your preferences between attribute lists are called **weakly separable**. So weak separability means that your preference between two attribute lists that differ only in one position does not depend on the attributes in the other positions.

Consider the following preferences between four possible flats.

$$\langle 50\text{m}^2, \text{central}, £500 \rangle \succ \langle 40\text{m}^2, \text{beach}, £500 \rangle$$
$$\langle 40\text{m}^2, \text{beach}, £400 \rangle \succ \langle 50\text{m}^2, \text{central}, £400 \rangle$$

Among flats that cost £500, you prefer central 50 m^2 flats to 40 m^2 flats at the beach. But among flats that cost £400, your preferences are reversed: you prefer 40 m^2 beach flats to 50 m^2 central flats. In a sense, your preferences for size and location depend on price. Nonetheless, your preferences may well be weakly separable.

That's why weak separability is called 'weak'. To rule out the present kind of dependence, we need to strengthen the concept of separability. Your preferences are **strongly separable** if your ranking of lists that differ in *one or more positions* does not depend on the attributes in the remaining positions. In the example, your ranking of $\langle 50\text{m}^2, \text{central}, - \rangle$ and $\langle 40\text{m}^2, \text{beach}, - \rangle$ depends on how the blank ('—') is filled in, and so your preferences aren't strongly separable.

Exercise 7.3 ⋆⋆

Suppose all you care about is the degree of pleasure of you and your three friends. And suppose you prefer states in which you four experience equal pleasure to states in which your degrees of pleasure are very different. For example,

you prefer $\langle 2,2,2,2\rangle$ to $\langle 2,2,2,8\rangle$, and you prefer $\langle 8,8,8,8\rangle$ to $\langle 8,8,8,2\rangle$. Are your preferences weakly separable? Are they strongly separable?

Exercise 7.4 ⋆⋆

Which of the following preferences violate weak separability or strong separability, based on the information provided?

a.	b.	c.
$\langle A_1,B_1,C_3\rangle \succ \langle A_3,B_1,C_1\rangle$	$\langle A_1,B_3,C_1\rangle \succ \langle A_1,B_3,C_2\rangle$	$\langle A_1,B_3,C_2\rangle \succ \langle A_1,B_1,C_2\rangle$
$\langle A_3,B_2,C_1\rangle \succ \langle A_1,B_2,C_3\rangle$	$\langle A_1,B_2,C_2\rangle \succ \langle A_1,B_2,C_3\rangle$	$\langle A_2,B_3,C_2\rangle \succ \langle A_2,B_1,C_2\rangle$
$\langle A_3,B_2,C_3\rangle \succ \langle A_3,B_2,C_1\rangle$	$\langle A_3,B_2,C_3\rangle \succ \langle A_3,B_1,C_3\rangle$	$\langle A_1,B_1,C_1\rangle \succ \langle A_1,B_3,C_1\rangle$

In 1960, Gérard Debreu proved that strong separability is exactly what is needed to ensure additivity.

To state Debreu's result, let's say that an agent's preferences over attribute lists have an **additive representation** if there is a value function V assigning numbers to the lists and there are subvalue functions $V_1, V_2, \ldots, V_n$ assigning numbers to attributes in the individual positions of the lists such that the following two conditions are satisfied. First, the preferences are represented by V. That is, for any two lists A and B,

$$A \succ B \text{ iff } V(A) > V(B), \text{ and } A \sim B \text{ iff } V(A) = V(B).$$

Second, the value assigned to any list $\langle A_1, A_2, \ldots, A_n\rangle$ equals the sum of the subvalues assigned to the items on the list:

$$V(\langle A_1, A_2, \ldots, A_n\rangle) = V_1(A_1) + V_2(A_2) + \ldots + V_n(A_n).$$

Now, in essence, Debreu's theorem states that if preferences over attribute lists are complete and transitive, then they have an additive representation if and only if they are strongly separable.

A technical further condition is needed if the number of attribute combinations is uncountably infinite; we'll ignore that. Curiously, the result also requires that there are at least three attributes that matter to the agent. For two attributes, a different condition called 'double-cancellation' is required instead of separability. Double-cancellation says that if $\langle A_1,B_1\rangle \succsim \langle A_2,B_2\rangle$ and $\langle A_2,B_3\rangle \succsim \langle A_3,B_1\rangle$ then $\langle A_2,B_3\rangle \succsim \langle A_3,B_2\rangle$. But let's just focus on cases with at least three relevant attributes.

One consequence of Debreu's theorem may also be worth noting: if the agent's preferences are defined over a sufficiently rich set of possibilities, then the value function V that additively represents the preferences is unique except for the choice of unit and zero. Additivity therefore opens the way to another potential response to the ordinalist challenge. The ordinalists claimed that utility assignments are arbitrary as long as they respect the agent's preference order. In response, one might argue that strongly separable preferences should be represented by an additive utility (or value) function. The utilities representing strongly separable preferences would then have an interval scale.

Exercise 7.5 ⋆⋆⋆

Show that whenever V additively represents an agent's preferences, then so does any function V' that differs from V only by the choice of zero and unit. That is, assume that V additively represents an agent's preferences, so that for some subvalue functions $V_1, V_2, \ldots, V_n$,

$$V(\langle A_1, A_2, \ldots, A_n \rangle) = V_1(A_1) + V_2(A_2) + \ldots + V_n(A_n).$$

Assume V' differs from V only by a different choice of unit and zero, which means that there are numbers $x > 0$ and y such that $V'(\langle A_1, A_2, \ldots, A_n \rangle) = x \cdot V(\langle A_1, A_2, \ldots, A_n \rangle) + y$. From these assumptions, show that there are subvalue functions $V'_1, V'_2, \ldots, V'_n$ such that

$$V'(\langle A_1, A_2, \ldots, A_n \rangle) = V'_1(A_1) + V'_2(A_2) + \ldots + V'_n(A_n).$$

Exercise 7.6 ⋆⋆

Imagine you can freely choose four courses for next semester. You assess each course by a range of criteria (such as whether the course will teach you anything useful). On this basis, you determine an overall ranking of the courses and sign up for the top four. Why might this not be a good idea?

7.4 Separability across time

According to psychological hedonism, the only thing people ultimately care about is their personal pleasure. But pleasure isn't constant. So the hedonist conjecture leaves open how people rank different ways pleasure can be distributed over a lifetime. Unless an agent just cares about her pleasure at a single point in time, a basic desire for pleasure is really a concern for a lot of things: pleasure now, pleasure tomorrow, pleasure the day after, and so on. We can think of these as the "attributes" in the agent's value function. The hedonist's value function somehow aggregates the value of pleasure experienced at different times.

To keep things simple, let's pretend that pleasure does not vary within any given day. We might then model a hedonist value function as a function that assigns numbers to lists like $\langle 1, 10, -1, 2, \dots \rangle$, where the elements in the list specify the agent's degree of pleasure today (1), tomorrow (10), the day after (-1), and so on. Such attribute lists in which successive positions correspond to successive points in time are called **time streams**.

A hedonist agent would plausibly prefer more pleasure to less at any point in time, no matter how much pleasure there is before or afterwards. If so, their preferences between time streams are weakly separable. Strong separability is also plausible: whether the agent prefers a certain amount of pleasure on some days to a different amount of pleasure on these days should not depend on how much pleasure the agent has on other days. It follows by Debreu's theorem that the value the agent assigns to a time stream can be determined as the sum of the subvalues she assign to the individual parts of the stream. That is, if p_1, p_2, …, p_n are the agent's degrees of pleasure on days $1, 2, \dots, n$ respectively, then there are subvalue functions $V_1, V_2, \dots, V_n$ such that

$$V(\langle p_1, p_2, \dots, p_n \rangle) = V_1(p_1) + V_2(p_2) + \dots + V_n(p_n).$$

We can say more if we make one further assumption. Suppose an agent prefers stream $\langle p_1, p_2, \dots, p_n \rangle$ to an alternative $\langle p'_1, p'_2, \dots, p'_n \rangle$. Now consider the same streams with all entries pushed one day into the future, and prefixed with the same degree of pleasure p_0. So the first stream turns into $\langle p_0, p_1, p_2, \dots, p_n \rangle$ and the second into $\langle p_0, p'_1, p'_2, \dots, p'_n \rangle$. Will the agent prefer the modified first stream to the modified second stream, given that she preferred the original first stream? If the answer is yes, then her preferences are called **stationary**. From a hedonist perspective, stationarity seems plausible: if there's more aggregated pleasure

in $\langle p_1, p_2, \ldots, p_n \rangle$ than in $\langle p'_1, p'_2, \ldots, p'_n \rangle$, then there is also more pleasure in $\langle p_0, p_1, p_2, \ldots, p_n \rangle$ than in $\langle p_0, p'_1, p'_2, \ldots, p'_n \rangle$.

It is not hard to show that if preferences over time streams are separable and stationary (as well as transitive and complete), then they can be represented by a value function of the form

$$V(\langle A_1, \ldots, A_n \rangle) = V_1(A_1) + \delta \cdot V_1(A_2) + \delta^2 \cdot V_1(A_3) \ldots + \delta^{n-1} \cdot V_1(A_n),$$

where δ is some number. The interesting thing here is that the subvalue function for all times equals the subvalue function V_1 for the first time, scaled by the exponential **discounting factor** δ^i.

So if a hedonist agent has strongly separable and stationary preferences, then the only remaining question is to what extent she discounts future pleasure. If $\delta = 1$, she values pleasure equally no matter when it occurs. If $\delta = 1/2$, then one unit of pleasure today is worth twice as much as one unit of pleasure tomorrow, four times as much as one unit of pleasure the day after tomorrow, and so on.

Exercise 7.7 ⋆

Consider the following streams of pleasure:

S1: $\langle 1, 2, 3, 4, 5, 6, 7, 8, 9 \rangle$
S2: $\langle 9, 8, 7, 6, 5, 4, 3, 2, 1 \rangle$
S3: $\langle 1, 9, 2, 8, 3, 7, 4, 6, 5 \rangle$
S4: $\langle 9, 1, 8, 2, 7, 3, 6, 4, 5 \rangle$
S5: $\langle 5, 5, 5, 5, 5, 5, 5, 5, 5 \rangle$

Assuming present pleasure is valued in proportion to its degree, so that $V_1(p) = p$ for all degrees of pleasure p, how would a hedonist agent with separable and stationary preferences rank these streams, provided that (a) $\delta = 1$, (b) $\delta < 1$, (c) $\delta > 1$? (You need to give three answers.)

Even if you're not a hedonist, you probably care about some things that can occur (and re-occur) at different times: talking to friends, going to concerts, having a glass of wine, etc. The formal results still apply. If your preferences over the relevant time streams are separable and stationary, then they are fixed by your subvalue function for having the relevant events (talking to friends, etc.) right now and a discounting parameter δ.

Some have argued that stationarity and separability across times are requirements of rationality. Some have even suggested that the only rationally defensible

discounting factor is 1, on the ground that we should be impartial with respect to different parts of our life.

One argument in favour of stationarity is that – under certain modelling assumptions – it is required to protect the agent from a kind of disagreement with her future self. To illustrate, suppose you prefer getting £100 now to getting £105 tomorrow, but you also prefer £105 in 11 days over £100 in 10 days. These preferences violate stationarity. For if you prefer $\langle £100, £0, \dots \rangle$ to $\langle £0, £105, \dots \rangle$ (the entries in the positions specifying how much money you get on successive days), then by stationarity you also prefer $\langle £0, £100, £0, \dots \rangle$ to $\langle £0, £0, £105, \dots \rangle$, and $\langle £0, £0, £100, £0, \dots \rangle$ to $\langle £0, £0, £0, £105, \dots \rangle$, and so on; so £100 in 10 days should be preferred to £105 in 11 days. Now suppose your (non-stationary) preferences remain the same for the next 10 days. At the end of this time, you then still prefer £100 now over £105 tomorrow. But your new "now" is your old "in 10 days". So your new preferences disagree with those of your earlier self in the sense that what you now regard as better is what your earlier self regarded as worse. That kind of disagreement is called **time inconsistency**.

Empirical studies show that time inconsistency is very common, and many instances of it can certainly appear problematic. For example, people often prefer their future selves to study, eat well, and exercise, but choose burgers and TV for today.

On the other hand, many violations of stationarity and even separability across time look perfectly sensible. For example, suppose you value having a glass of wine every now and then. But only now and then; you don't want to have wine every day. It follows that your preferences violate both separability and stationarity. You violate stationarity because even though you might prefer a stream $\langle \text{wine}, \text{no wine}, \text{no wine}, \dots \rangle$ to $\langle \text{no wine}, \text{no wine}, \text{no wine}, \dots \rangle$, your preference reverses if both streams are prefixed with wine (or many instances of wine). You violate separability because whether you regard having wine in n days as desirable depends on whether you will have wine right before or after these days.

Even if an agent only cares about pleasure, it is not obvious why a rational agent might not (say) prefer relatively constant levels or pleasure over wildly fluctuating levels, or the other way round. Either preference would violate both stationarity and separability.

On standard ways of modelling preferences over time streams (which ignores what happened in the past), these preferences are time inconsistent. But that kind of time inconsistency does not look problematic.

What shows up in the more problematic kind of time inconsistency concerning studying, food, or exercise, is that preferences have a range of different sources, as I emphasized in chapter 5. When we reflect on having fries or salad now, we are more influenced by spontaneous cravings than when we consider the same options in the distant future.

If different sources or kinds of preference pertain to different aspects of the world, then the results we have reviewed can also clarify how these sources are aggregated into the agent's all-things-considered preference. In particular, if the agent's preferences are separable across the relevant aspects, then (and only then) the all-things-considered preferences can be understood to result by adding up (and possibly scaling) independent scores assigned by the different sources.

7.5 Separability across states

Let's briefly return to decision problems. In a decision problem, every available act leads to a particular outcome in each of the relevant states. Standard decision theory assumes that a rational agent prefers an act A to an act B in a given decision problem just in case A has greater expected utility, defined as

$$EU(A) = U(O_1)\mathrm{Cr}(S_1) + U(O_2)\mathrm{Cr}(S_2) + \ldots + U(O_n)\mathrm{Cr}(S_n),$$

where $S_1, S_2, \ldots, S_n$ are the states and $O_1, O_2, \ldots, O_n$ are the various outcomes of act A in those states.

Standard decision theory therefore assumes that the only thing that matters to the agent's preferences between acts, in any fixed decision problem, are the possible outcomes. If two acts lead to the same outcomes in all states, the agent will be indifferent between them. We can therefore model the agent's preferences between acts as preferences between lists of outcomes, one for each state. For example, in the mushroom problem from chapter 1, eating the mushroom can be modelled as $\langle \text{satisfied}, \text{dead} \rangle$, and not eating as $\langle \text{hungry}, \text{hungry} \rangle$.

Now, if an agent ranks acts by their expected utility, then her preferences between acts have an additive representation, since they are represented by a function V whose values are determined by adding up subvalues assigned to the individual outcomes: the function V is the EU function; the subvalue assigned to outcome O_1 is $U(O_1)\mathrm{Cr}(S_1)$, and so on.

By Debreu's theorem, rational preferences have an additive representation if and only if they are strongly separable. So standard decision theory implies that

preferences between acts are (strongly) separable across states, meaning that the desirability of an act's outcome in one state does not depend on the outcomes in other states.

Admittedly, this is an elaborate path to a fairly obvious result. I mention it for two reasons. First, it shows that the two responses to the ordinalist challenge are actually closely related. In effect, Ramsey, Savage, and von Neumann and Morgenstern assume that rational preferences are separable across states, and that these preferences should be represented additively, in terms of expected utilities.

Second, a general consequence of separability is that the relevant preferences are insensitive to certain "shapes" in the distribution of subvalues. In particular, separable preferences cannot prefer even distributions to uneven distributions. That points at a potential problem with the MEU Principle and any other decision rule that implies separability across states. For example, consider the following schematic decision problem:

	State 1 ($1/2$)	State 2 ($1/2$)
A	Outcome 1 (+10)	Outcome 1 (+10)
B	Outcome 2 (-10)	Outcome 3 (+30)

Option A leads to a guaranteed outcome with utility 10, while option B leads either to a much better outcome or to a much worse one. The expected utilities are the same, but one might think an agent might rationally prefer the safe option A just because it is safe – because the utility distribution $\langle 10, 10 \rangle$ is more even than $\langle \text{-}10, 30 \rangle$. Much more on that in the next chapter.

Exercise 7.8 ⋆⋆

Where in their axioms do Savage and von Neumann and Morgenstern postulate a kind of separability across states?

7.6 Harsanyi's "proof of utilitarianism"

The ordinalist movement, which rejected the concept of utility as a well-defined numerical quantity, posed a challenge not only to the MEU Principle, but also to utilitarianism in ethics. According to utilitarianism, an act is right just in case it brings about the best available state of the world; a state of the world is better than an alternative just in case the sum of the utility of all people in that state

is greater than in the alternative. Without a numerical (and not just ordinal) measure of utility, the second of these claims becomes meaningless. We would need a new criterion for ranking states of the world.

One such criterion was proposed by Pareto. Recall that Pareto did not deny that people have preferences. So if we want to rank two states of the world, we can meaningfully ask which of them people prefer. And that allows us to define at least a partial order on the possible states:

The Pareto Condition

If everyone is indifferent between A and B, then A and B are equally good; if at least one person prefers A to B and no one prefers B to A, then A is better than B.

Unlike classical utilitarianism, however, the Pareto Condition offers little moral guidance. For instance, while classical utilitarianism suggests that one should harvest the organs of an innocent person in order to save ten others, the Pareto Condition does not settle whether it would be better or worse to harvest the organs, given that the person to be sacrificed ranks the options differently than those who would be saved.

Exercise 7.9 (The Condorcet Paradox) ⋆

A "democratic" strengthening of the Pareto condition might say that whenever *a majority* of people prefer A to B, then A is better than B. But consider the following scenario. There are three relevant states: A, B, C, and three people. Person 1 prefers A to B to C. Person 2 prefers B to C to A. Person 3 prefers C to A to B. If betterness is decided by majority vote, is A better than B? How about C and A, and B and C?

In 1955, John Harsanyi proved a remarkable theorem that seemed to rescue, and indeed vindicate, classical utilitarianism.

To begin, Harsanyi assumes that there is a betterness order between states of the world which is also defined for lotteries between such states. That is not yet a substantive premise, as we have not yet made any substantive assumptions about the order.

Harsanyi's first premise is that the order satisfies the axioms of von Neumann and Morgenstern. By von Neumann and Morgenstern's representation theorem, it follows that the betterness order is represented by a ("social") utility function that is unique except for the choice of unit and zero.

Second, Harsanyi assumes that the betterness order satisfies the Pareto condition (both for states and for lotteries).

Finally, Harsanyi assumes that each person – of which he assumes for simplicity that there is a fixed number n – has personal preferences between the relevant states and lotteries, and that these preferences also satisfy the von Neumann and Morgenstern axioms. So they are represented by n personal utility functions.

Note that the Pareto condition states a simple kind of separability across people. The assumption that social and personal utility rank lotteries by their expected utility, which follows from the von Neumann and Morgenstern construction, amounts to separability in another dimension, across states. As it turns out, Debreu's results can be strengthened for cases in which the attributes are separable across two independent dimensions (here, people and states). Drawing on this result, Harsanyi showed that it follows from the above three assumptions that the individual and social preferences are represented by utility functions U_s and $U_1, U_2, \ldots, U_n$ such that the social utility function is simply the sum of the individual utility functions: for any state or lottery A,

$$U_s(A) = U_1(A) + U_2(A) + \ldots + U_n(A).$$

And that looks just like classical utilitarianism.

On closer inspection, things are less clear-cut. For a start, recall that the utility functions established by von Neumann and Morgenstern's representation theorem have arbitrary units and zeroes. So if according to one adequate representation of our preferences, my utility for a given state is 10 and yours is 0, then according to another equally adequate representation, my utility for the state is 10000 and yours -3. All Harsanyi's theorem tells us is that there is *some* utility representation of our individual preferences relative to which our utilities add up to social utility. This is compatible with the assumption that social utility is almost entirely determined by the preferences of a single person, because her utilities are scaled so as to dwarf all the others. That does not look like classical utilitarianism.

Also, anyone who is not already a utilitarian should probably reject the Pareto Condition. After all, the condition implies that the only thing that matters, from a moral perspective, is the satisfaction of people's preferences. If anything else had any moral weight – whether people's rights are respected, whether animals suffer, whether God's commands are obeyed, or whatever – then it could happen that everyone is indifferent between A and B, and yet A is actually better.

In general, if someone seems to offer a mathematical proof of a substantive nor-

mative principle, you can be sure that either the principle isn't really established or it has been smuggled in through the premises.

7.7 Further reading

For an opinionated review of various positions on time (in)consistency, see

- Tomasz Żuradzki: "Time-biases and Rationality: The Philosophical Perspectives on Empirical Research about Time Preference" (2016)

A lucid introduction to Harsanyi's argument for utilitarianism is

- John Broome: "Utilitarianism and Expected Utility" (1987)

Most of the remaining topics in this chapter covered elementary ideas from a subject known as "multi-attribute utility theory". I don't know any introduction to this subject that is reasonably short and good.

Essay Question 7.1

Is time inconsistency always irrational? Can you explain why, or why not?

Essay Question 7.2

Following up on a thought at the end of section 7.4: Can you think of a way to define separability directly for sources of utility, without assuming that different kinds of motives pertain to different aspects of the world?

8 Risk

8.1 Why maximize expected utility?

So far, we have largely taken for granted that rational agents maximize expected utility. It is time to put this assumption under scrutiny.

In chapter 1, I motivated the MEU Principle by arguing that an adequate decision rule should consider all the outcomes an act might bring about – not just the best, the worst, or the most likely – and that it should weigh outcomes in proportion to their probability, so that more likely outcomes are given proportionally greater weight. These are fairly natural assumptions, and they rule out many alternatives to the MEU Principle.

We encountered another, quite different, argument for the MEU Principle in chapter 6. The argument began with the assumption that utility can be measured by the methods described by von Neumann and Morgenstern or Savage. The MEU Principle can then be shown to reduce to certain "axioms" concerning the agent's preferences. To conclude the argument, we would have to convince ourselves that these axioms are genuine norms of rationality. We will look at some apparent counterexamples below.

Yet another argument for the MEU Principle was hiding in chapter 5. There we saw that the desirability (or utility) of a proposition can plausibly be understood as a probability-weighted average of the desirability of its parts, as described by Jeffrey's axiom. Now consider a schematic decision problem with two acts and two states.

	S_1	S_2
A	O_1	O_2
B	O_3	O_4

Choosing A effectively means choosing to bring about $O_1 \vee O_2$; choosing B means choosing to bring about $O_3 \vee O_4$. Let's assume the four outcomes are logically incompatible with each other. (We can always make them incompatible by

adding more information. For example, if O_1 and O_2 are not incompatible, we can redefine them as $O_1 \wedge S_1$ and $O_2 \wedge S_2$, respectively.)

By Jeffrey's axiom,

$$U(O_1 \vee O_2) = U(O_1) \cdot \mathrm{Cr}(O_1/O_1 \vee O_2) + U(O_2) \cdot \mathrm{Cr}(O_2/O_1 \vee O_2).$$

Since choosing A is certain to lead to either O_1 or O_2, and choosing B is certain to lead to O_3 or O_4, we also have

$$\mathrm{Cr}(O_1 \vee O_2) = \mathrm{Cr}(A).$$

Moreover, on the supposition that $O_1 \vee O_2$ is true, it is certain that O_1 comes about just in case state S_1 obtains:

$$\mathrm{Cr}(O_1/O_1 \vee O_2) = \mathrm{Cr}(S_1/O_1 \vee O_2).$$

Together, the previous two observations entail that

$$\mathrm{Cr}(O_1/O_1 \vee O_2) = \mathrm{Cr}(S_1/A).$$

But in a well-defined decision matrix, the states must be independent of the acts. So it looks like we can simplify further:

$$\mathrm{Cr}(O_1/O_1 \vee O_2) = \mathrm{Cr}(S_1).$$

By the same reasoning, $\mathrm{Cr}(O_2/O_1 \vee O_2) = \mathrm{Cr}(S_2)$. The above instance of Jeffrey's axiom can therefore be rewritten as follows:

$$U(A) = U(O_1) \cdot \mathrm{Cr}(S_1) + U(O_2) \cdot \mathrm{Cr}(S_2).$$

This says that the *expected utility* of act A equals the *utility* of A! Since utility is defined as a measure of desirability, the MEU principle therefore says that rational agents prefer to bring about more desirable propositions rather than less desirable propositions. (We will reconsider the present argument in chapter 9.)

We have seen three arguments in favour of the MEU Principle. I will mention two more, before I turn to objections.

The next argument is simply that the MEU Principle seems to deliver the correct result in many concrete situations. This includes situations in which it is obvious what one should do (like in the mushroom problem from page p.9), but also situations in which many people are tempted to give a different answer.

For example, people often fail to give appropriate weight to low-probability outcomes with very high or very low utility. Many people worry about dying in a plane crash or a terrorist attack, and take steps to avoid these events, but don't think twice about driving to the mall, even after being informed that they are more likely to die on their way to the mall than on a plane trip or in a terror attack. This violates the norms of Bayesian rationality, and on reflection it does seem irrational.

Exercise 8.1 ⋆

In the National Lottery, a £2 ticket typically has an expected payoff of around £1. Many people are aware of this fact but still play the lottery. One explanation is that they are violating the MEU Principle. Can you think of an alternative explanation?

Another example. Suppose you try to avoid plane trips not because you are afraid of a crash, but because you care about your carbon footprint. Your friend argues that your behaviour is irrational because avoiding plane trips actually won't affect any carbon emissions: no plane is going to stay on the ground just because you don't fly. Is she right?

It is certainly unlikely that fewer flights will be scheduled as a result of a single person deciding not to go on a plane. On the other hand, the number of flights is sensitive to demand: if, one by one, fewer people decide to fly, at some point fewer flights will be scheduled. So there must be some chance that avoiding a single plane trip will reduce overall air traffic. To be sure, the chance is low. On the other hand, the reduction in carbon emissions would be high. One can estimate the *expected* reduction in carbon emissions: the probability-weighted average of the reduction in carbon emissions. Depending on the flight route, it typically works out to be a little less than the flight's emissions divided by the number of seats on the plane. This is how much overall carbon emissions are reduced, on average, as a result of one passenger deciding not to take a flight. And so it makes perfect sense to avoid plane trips if you want to reduce your carbon footprint. Your friend is wrong.

Even Nobel-price winning decision theorists are not immune to this kind of error. In 1980, John Harsanyi published a paper in which he argued that utilitarian citizens who care a lot about the common good would still have little incentive to participate in elections, given that any individual vote is almost certain not to make a difference. Here is one of Harsanyi's simplified examples.

Example 8.1

"1000 voters have to decide the fate of a socially very desirable policy measure M. All of them favor the measure. Yet it will pass only if all 1000 voters actually come to the polls and vote for it. But voting entails some minor costs in terms of time and inconvenience. The voters cannot communicate and cannot find out how many other voters actually voted or will vote."

Harsanyi claims that even if the 1000 eligible voters are utilitarians whose aim is to bring about the best overall result for everyone, defeat of the measure is likely, since "each voter will vote only if he is reasonably sure that all other 999 voters will vote".

Harsanyi is making the same mistake as your hypothetical friend above. To flesh out the scenario, let's assume that each voter would lose 1 degree of pleasure by voting. For the case to be interesting, we can assume that the very desirable measure M would more than offset the cost of voting, so that the group of 1000 voters is better off if everyone votes than if everyone stays at home. Since the total degree of pleasure in the society is reduced by 1000 if everyone votes, the measure M must therefore increase the total degree of pleasure in the society by more than 1000. Now consider a utilitarian voter who values outcomes by their effect on the total degree of pleasure in the society. If the agent is certain that only, say, 900 others will vote, it would clearly be sensible for her (in line with the MEU Principle) to stay at home. But she doesn't need to be "reasonably sure", as Harsanyi claims, that all other 999 will vote, to make it worthwhile for her to go out and vote – just as you don't need to be reasonably sure that fewer flights will be scheduled if you don't take a plane. If you do the math, you can see that voting maximizes expected utility even if the probability of all others voting is as low as 0.001.

Exercise 8.2 ⋆⋆⋆

Do the math. That is, describe the decision matrix for a voter in Harsanyi's scenario, and confirm that voting maximizes expected utility if the probability of all others voting is 0.001.

8.2 The long run

The last argument for the MEU Principle that I want to mention looks at the long-term consequences of maximizing expected utility.

Suppose you repeatedly toss a fair coin, keeping track of the total number of heads and tails. You will find that over time, the proportion of each outcome approaches its objective probability, $1/2$. After one toss, you will have 100% heads or 100% tails. After ten tosses, it's very unlikely that you'll still have 100% heads or 100% tails. But 60% heads and 40% tails wouldn't be unusual. The probability of getting 40% tails or less in 10 independent tosses of a coin is 0.377. For 100 tosses, it is 0.028; for 1000, it is less than 0.000001. After 1000 tosses, the probability that the proportion of tails lies between 45% and 55% is 0.999.

In general, the probability axioms entail that if there is a sequence of "trials" $T_1, T_2, T_3 \ldots$ with the same possible outcomes (like heads and tails) and the same probabilities for the outcomes, independent of earlier outcomes, then the probability that the *proportion* of any outcome in the sequence differs from its *probability* by more than an arbitrarily small amount ϵ converges to 0 as the number of trials gets larger and larger. This is known as the **law of large numbers**. Loosely speaking: in the long run, probabilities turn into proportions.

How is that relevant to the MEU Principle? Well, consider a bet on a fair coin flip: if the coin lands heads, you get £1, otherwise you get £0. The bet costs £0.40. If you are offered this deal again and again, the law of large numbers entails that the percentage of heads will (with high probability) converge to 50%. So if you buy the bet each time, you can be confident that you will loose £0.40 in about half the trials and win £0.60 in the other half. The £0.10 *expected payoff* turns into the *average payoff*. In this kind of scenario, the MEU Principle effectively says that you should prefer acts with greater average utility over acts with lower average utility. Which seems obviously correct.

In reality, of course, there are limits to how often one can encounter the very same decision problem. "In the long run, we are all dead", as John Maynard Keynes quipped. The practical relevance of the above argument therefore derives not just from the law of large numbers, but from the fact that probabilities can be expected to converge to proportions *reasonably fast*: it does not take millions of tosses until the percentage of heads is almost certain to exceed 40%.

The argument also assumes that the same decision problem is faced over and over. In practice, this rarely happens. But the argument can be generalised. Suppose an agent faces a sequence of n decision problems, which may involve

different outcomes, different states, and different probabilities. One can show that if the states in the various problems are probabilistically independent and the relevant probabilities and utilities are not too extreme, then the *average* utility is still likely to converge to the *expected* utility, and it will do so relatively fast.

From what I said, you might expect that professional gamblers and investors generally put their money on the options with greatest expected payoff, which would give them the greatest overall profit in the long run. But they do not. – Those who do don't remain professional gamblers or investors for long. To see why, imagine you are offered to invest in a startup company that attempts to find a cure for snoring. If the company succeeds, your investment will pay back tenfold. If they don't, the investment is lost. The chance of success is 20%, so the expected return is $0.2 \cdot 1000\% + 0.8 \cdot 0\% = 200\%$. Even if this exceeds the expected return of all other investment possibilities, you would be mad to put all your money into that company. If you repeatedly face that kind of decision and go all-in each time, then after ten rounds you are bankrupt with a probability of $1 - 0.2^{10} = 0.9999998976$.

This does not contradict the law of large numbers. In the startup example, you are not facing the same decision problem again and again. If you lose all your money in the first round, you don't have anything left to invest in later rounds. Still, the example shows that maximizing expected utility does not always mean maximizing long-term actual utility. More importantly, it suggests that there is something wrong with the MEU Principle. Sensible investors balance expected returns and risks. A safe investment with lower expected returns is often preferred to a risky investment with greater expected returns. Shouldn't we adjust the MEU Principle, so that agents can factor in risk in addition to expected utility?

Exercise 8.3 ⋆⋆

Every year, an investor is given £100,000, which she can invest either in a risky startup of the kind described (a different one each year), or put in a bank account at 0% interest. If she always chooses the second option, she will have £1,000,000 after ten years.

(a) What are the chances that she would do at least as well (after ten years) if she always chooses the first option, without reinvesting previous profits? (Hint: Compute the chance that she would do worse.)

(b) How does the answer to (a) mesh with my claim in the text that an investor who always goes with the risky option is virtually guaranteed to go bankrupt?

8.3 Risk aversion

Aversion to risk is common, and does not seem irrational. Let's see if it poses a genuine threat to the MEU Principle.

The standard way to measure risk aversion is to offer people gambles. Consider a lottery with an 80% chance of £0 and a 20% chance of £1000. The lottery's expected payoff is £200. Given a choice between the lottery and £100 for sure, a risk averse agent might prefer the £100. Can we account for these preferences, without giving up the MEU Principle?

Yes. We only need to assume that, for this agent, the difference in utility between £1000 and £100 is less than five times the difference in utility between £100 and £0. For example, if $U(£0) = 0$, $U(£100) = 1$, and $U(£1000) = 4$, then the lottery has expected utility $0.8 \cdot 0 + 0.2 \cdot 4 = 0.8$, which is less than the guaranteed utility of the £100.

Along these lines, the standard model of risk aversion in economics is to assume that utility is a "concave function of money", meaning that the amount of utility an extra £100 would add to an outcome of £1000 is less than the amount of utility the same £100 would add to an lesser outcome of, say, £100. We have already encountered this phenomenon in chapter 5, where we saw that for most people, money has declining marginal utility: the more you have, the less utility you get from an extra £100. According to standard economics, risk aversion is the flip side of declining marginal utility.

This should seem strange. Intuitively, the fact that the same amount of money becomes less useful the more money you already have has nothing to do with risk. Money could have declining marginal utility even for an agent who loves risk. Conversely, an agent might value every penny as much as the previous one, but shy away from risks.

The problem is that risk neutrality combined with declining marginal utility seems to support the exact same choices as risk aversion combined with non-declining marginal utility. So if an agent's utility function is defined through her preferences or choice dispositions over monetary gambles – by the von Neumann

and Morgenstern method, perhaps – then the intuitive difference between risk aversion and declining marginal utility collapses.

You can spin this both ways. You can conclude that we should reject the intuitive difference between risk aversion and declining marginal utility. Or you can hold on to the intuitive difference and conclude that utilities cannot be defined through preferences over monetary gambles.

In any case, things are not so simple. The following scenario, due to Maurice Allais, shows that risk aversion is not equivalent to declining marginal utility after all. It also appears to show that risk aversion is incompatible with the MEU Principle.

Example 8.2 (Allais's Paradox)

A ball is drawn from an urn containing 80 red balls, 19 green balls, and 1 blue ball. Consider first a choice between the following two lotteries. Which do you prefer?

	Red	Green	Blue
A	£0	£1000	£1000
B	£0	£1200	£0

Next, consider the alternative lotteries *C* and *D*, based on the same draw from the urn. Which of these do you prefer?

	Red	Green	Blue
C	£1000	£1000	£1000
D	£1000	£1200	£0

The MEU Principle does not settle the answer to either question. But it seems to rule out a combination of preferences many people in fact express, namely a preference for *B* over *A*, and for *C* over *D*.

Why might you have these preferences? Notice that the second choice is essentially one between £1000 for sure and a gamble in which you might get either £1000 (most likely) or £0 (least likely) or £1200. If you're risk averse, it makes sense to take the sure £1000. By contrast, in the first choice the most likely outcome is £0 either way, and it seems reasonable to take the 19% chance of getting £1200 rather than the 20% chance of getting £1000.

You can easily check that there is no way of assigning utilities to monetary

payoffs that makes these preferences conform to the MEU Principle. In fact, the preference for B over A and C over B appears to violate the Independence axiom both in von Neumann and Morgenstern's form and in Savage's form. To see this (for the von Neumann and Morgenstern version), let L_1 be a lottery that pays £1000 if a green or blue ball is drawn (otherwise £0), and let L_2 be a lottery that pays £1200 if a green ball is drawn (otherwise £0). Note that A is (in effect) a lottery that leads to L_1 with probability 0.2 and otherwise to £0. Similarly, B is a lottery that leads to L_2 with probability 0.2 and otherwise to £0. By the Independence axiom, if $L_1 \succsim L_2$, then $A \succsim B$. By parallel reasoning with C and D, it follows from Independence that

1. if $L_1 \succsim L_2$, then $A \succsim B$ and $C \succsim D$; and
2. if $L_2 \succsim L_1$, then $B \succsim A$ and $D \succsim C$.

No matter how you rank L_1 and L_2, you can't have $B \succ A$ and $C \succ D$.

Exercise 8.4 ⋆⋆

Does the preference of B to A and C to D violate strong separability, weak separability, or neither? (Explain briefly.)

Now, the fact that many people prefer B to A and C to D in Allais's Paradox is, by itself, not a problem for the MEU Principle. After all, these people may simply be irrational. The problem is that the Allais preferences seem to make perfect sense for an agent who is risk averse.

One might respond that rational agents should not be averse to risk. But that would go against the Humean spirit of our model: we don't want to make substantive assumptions about what people should care about. Moreover, there are other problem cases in which this line of response is untenable. The following case goes back to Peter Diamond.

Example 8.3

A mother has a treat that she can give either to her daughter Abbie or to her son Ben. She considers three options: giving the treat to Abbie, giving it to Ben, and tossing a fair coin, so that Abbie gets the treat on heads and Ben on tails. Her decision problem might be summarized by the following matrix (assuming for simplicity that if the mother decides to give the treat directly to one of her children, she nonetheless tosses the coin, just for fun).

	Heads	Tails
Give treat to Abbie (A)	Abbie gets treat	Abbie gets treat
Give treat to Ben (B)	Ben gets treat	Ben gets treat
Let coin decide (C)	Abbie gets treat	Ben gets treat

The mother's preferences are $C \succ A$, $C \succ B$, $A \succ B$.

Like in Allais's Paradox, there is no way of assigning utilities to the outcomes that makes the mother's preferences conform to the MEU Principle. Yet these preferences are surely not irrational. The mother prefers C because C is the most fair of the three options. It would be absurd to claim that rational agents cannot value fairness.

Exercise 8.5 ⋆⋆⋆

Which of the von Neumann and Morgenstern axioms do the mother's preferences in example 8.3 appear to violate? (Explain briefly.)

8.4 Redescribing the outcomes

When confronted with an apparent counterexample to the MEU Principle, the first thing to check is always whether the decision matrix has been set up correctly. In particular, we need to check if the outcomes in the matrix specify all attributes that matter to the agent (and that vary between the outcomes).

The matrices in example 8.2 (Allais's Paradox) specify how much money you get depending on your choice and the drawn ball. But if you're genuinely risk averse, then you don't just care about how much money you will have. You also care about risk. So we need to add more information to the outcomes.

There are two ways of doing that. The first adds to the monetary payoffs further things that will happen as a result of the relevant choices and draws.

Consider the bottom right cell of the second matrix in example 8.2. What will happen if you chose D and the blue ball is drawn? You get £0. But you'd plausibly also feel frustrated about your bad luck: there was a 99% chance of getting at least £1000, and you got £0! You might also feel regret about your choice: if only you had chosen the safe alternative C, you'd now have £1000. You probably don't like feelings of frustration and regret. If so, these feelings should be added to the outcome. The outcome in the bottom right cell of the second matrix would then say something like '£0 and considerable frustration/regret'. By contrast,

consider the bottom right cell of the first matrix. If you choose B and the blue ball is drawn, you get £0. But the chance of getting £0 was 81%, so you'll be much less frustrated about your bad luck. You may still regret not having taken A, but since A was just as unsafe as B, you probably wouldn't think you made a terrible mistake. So the outcome in that cell might say something like '£0 and a little frustration/regret'. With these changes, the preference for B over A, and for C over D is easily reconciled with the MEU Principle.

Exercise 8.6 ⋆

Assign utilities to the outcomes in the two matrices, with the changes just described, so that $EU(B) > EU(A)$ and $EU(C) > EU(D)$.

The problem with this first type of response is that it doesn't always work. For example, suppose you face Allais's Paradox towards the end of your life. The ball will only be drawn after your death, and the money will go to your children. In that case, you will not be around to experience frustration or regret about the outcome, nor might your children, if the whole process is kept secret from them. But if you're risk averse, you might still prefer B to A and C to D.

The second strategy for redescribing outcomes gets around this problem. As before, we want to distinguish the outcomes in the bottom right cell of the two decision matrices. So let's ask again what will happen if you choose D and the blue ball is drawn. One thing that will happen is that you get £0. You may or may not experience frustration and regret. But here's another thing that is guaranteed to happen: you *will have chosen a risky option instead of a safe alternative*. If you are risk averse, then plausibly (indeed, obviously!) you care about whether your choices are risky. So we should put that into the outcome. By contrast, the outcome in the bottom right cell of the first matrix does not have this feature – that you will have chosen a risky option instead of a safe alternative – because the alternative A is not safe. So we can once again distinguish the two outcomes.

In general, the first strategy appeals only appeals to things that happen as a causal consequence of the relevant choice. Let's call such attributes **local**. So on the first strategy for redescribing outcomes, the outcomes must only involve local propositions whose truth is a causal consequence of the relevant act in the relevant state.

By contrast, the second strategy also allows for non-local attributes in the outcomes. To choose D rather than C *is* to choose a risky option instead of a safe alternative. That you chose a risky option is not a causal consequence of

your choice; it does not depend on the causal structure of the world; it is not a separate event that happens after your choice.

Let's say (following Lara Buchak) that outcomes are **individuated locally** if outcomes are only distinguished by features that are a causal consequence of the relevant act in the relevant state. If outcomes are not individuated locally, they are **individuated globally**.

Exercise 8.7 ⋆

Redescribe the outcomes in example 8.3 so that the mother's preferences conform to the MEU Principle.

Exercise 8.8 ⋆⋆

(a) In your solution to exercise 8.7, did you individuate the outcomes locally or globally?

(b) Either way, can you find another answer to the exercise that individuates outcomes the other way?

If global individuation of outcomes is allowed, we can defuse the threat of Allais's Paradox, even in cases without regret or frustration. Moreover, only this second strategy seems to capture what really motivates a risk averse agent. Intuitively, just as risk aversion is not the same as declining marginal utility of money, it is also not the same as fear of regret or frustration. If you are risk averse, then one of the things you care about is safety. Given that outcomes should specify everything the agent cares about – all her desires or motives – we should therefore include the safety or riskiness of a choice in its outcomes.

Nonetheless, the vast majority of decision theorists assume that outcomes must be individuated locally. The assumption is so common that it doesn't even have a name; let's call it **localism**. According to localism, in cases where local features like frustration or regret can't explain the Allais preferences, these preferences really do contradict the MEU Principle. We must either declare the preferences irrational or revise the MEU Principle.

Many such revisions to the MEU Principle have been proposed. An especially elegant recent example is Lara Buchak's *Risk-Weighted Expected Utility Theory*. On Buchak's model, the choiceworthiness of an act is not determined by the agent's credences and utilities, the latter of which pertain only to local attributes of outcomes. Instead, agents are also assumed to have a "risk function" representing

their attitudes towards the non-local features of safety and risk. Buchak shows that if an agent satisfies a variant of Savage's axioms with a weakened Independence Axiom, then they can be represented as maximizing risk-weighted expected utility.

But why should one go with localism in the first place? What is wrong with a global individuation of outcomes?

Exercise 8.9 ⋆⋆

Risk and fairness are two non-local attributes that many people care about. Can you think of another such attribute?

8.5 Localism

To understand the prevalence of localism, we need to go back in history. The formalism of expected utility theory was originally developed to compute the monetary value of gambles: a gamble with expected payoff £100 was assumed to be equivalent in value to £100. Bernoulli showed that this assumption is false, since money has declining marginal utility. But Bernoulli did not reject expected utility theory. Instead, he argued that we should distinguish between the monetary payoff itself and the utility of that payoff for an agent. A gamble is worth £100 if the gamble's expected *utility* equals the *utility* of £100.

For a while, people then understood utility to measure the degree of pleasure or welfare a choice might bring about. It was only in response to the ordinalist challenge of the early 20th century that utility was – officially – no longer defined as a measurable quantity causally downstream from the choice, but as reflecting the agent's motives or preferences before, or at the time of, the choice.

This second shift made it possible to regard expected utility theory as a general model of practical rationality, without assuming that rational agents only care about their personal pleasure or welfare.

But as we saw in chapter 5, the second shift never fully caught on. Many authors still assume that utility is a measure of the pleasure or welfare or wealth that results from a choice. On that conception, if an act A leads to a particular amount of pleasure/welfare/wealth in state S_1, and act B leads to the very same amount of pleasure/welfare/wealth in state S_2, then the outcome of choosing A in S_1 must have the same utility as the outcome of choosing B in S_2.

This is one explanation for the widespread acceptance of localism. If we were interested in modelling selfish agents who only care about their future pleasure

or welfare, localism would be harmless. But that is not our topic. We want to model agents who may care about other things – risk, for example, or fairness, or other people's welfare.

Even on the preference-based conception of utility, however, there is some pressure to endorse localism. Recall von Neumann's method for determining the utility of a reward c in comparison to other rewards a and b. The method considers the agent's preferences over certain lotteries between these rewards. For example, we would check whether she prefers a fair lottery between a and c to b. But if a, b, and c have non-local attributes, then these lotteries may well be logically impossible.

The problem is that anything that comes about as the result of a lottery inevitably has the non-local attribute of coming about as a result of that lottery. Suppose we wanted to use von Neumann's method to determine the utility function for the mother in example 8.3. Let a and b be the outcomes of directly giving the treat to Abby or Ben, respectively. If the mother cares about fairness, then one relevant (non-local) aspect of a and b is that who gets the treat is not decided by a chance process. By von Neumann's method, we should now ask whether the mother prefers some other outcome c to a *lottery between a and b*. This lottery would be a chance process that leads to outcomes which don't come about through a chance process. That's logically impossible.

Things are even worse for Savage. Savage assumes that for any possible outcomes a and b and any state S, the agent has preferences concerning prospects of the form $[S\,?\,a : b]$. But if the agent is allowed to care about arbitrary attributes, then one of these attributes might well be (or entail) S. And if, say, b entails S, then there can't be any prospect $[S\,?\,a : b]$, for such a prospect would lead to a if S and to b if $\neg S$. Savage's approach therefore requires a particular kind of localism, known as "state independence" of utilities.

The upshot is that if we want to allow agents to care about non-local attributes such as risk and fairness – as we should if we are interested in a general model of practical rationality – then we can't build that model on the foundations of von Neumann and Morgenstern or Savage (or Ramsey).

But we don't have to give up the idea that utilities (and credences) might be derived from preferences, or that the MEU Principle might be justified by more fundamental axioms concerning preferences. Other foundations can be provided. The best known of these alternatives was developed by Ethan Bolker and Richard Jeffrey in the 1960s.

In the Bolker/Jeffrey construction, the objects of utility are taken to be arbi-

trary propositions, not "rewards" or "outcomes" or lotteries between rewards or outcomes. No restrictions are imposed on an agent's basic desires. Like Savage, Bolker and Jeffrey show that if an agent's preferences between propositions satisfy certain axioms, then they can be represented by a probabilistic credence function Cr and a utility function U that satisfies Jeffrey's axiom – which, in turn, supports the MEU Principle by the argument outlined in section 8.1. I will not explain all the axioms or the details of the construction. However, I do want to mention that since preferences are now defined between arbitrary propositions, the gap between preferences and choice dispositions is even larger in the Bolker/Jeffrey construction than it was for von Neumann and Morgenstern or Savage.

This brings me to one final point in favour of localism. Localism promises to give the MEU Principle predictive power. Localism ensures that the very same outcome can figure in different decision problems. If outcomes can be individuated globally, then we can always find differences between outcomes in different decision problems. And then we can never find out, just from the agent's choices, that she violates transitivity, independence, or any of the other classical axioms. No matter what the agent does, there will always be some utility function relative to which all her choices maximize expected utility. Localism blocks this threat of trivialization.

But should we want to block the threat? Shouldn't we rather accept that on a broadly Humean conception of practical rationality, no pattern of behaviour could, all by itself, show that the agent is practically irrational? Couldn't the agent have a basic desire to display just this pattern of behaviour? It is not obvious that we should expect a general theory of practical rationality to make testable predictions, without any assumptions about the agent's beliefs and desires.

Localism in effect encodes one such assumption: that the agent only cares about local features of outcomes. But the assumption is not only implausible, it also barely helps with the trivialization worry. For even if we restrict ourselves to local properties, we can almost always find differences between outcomes in different decision problems. I suspect this explains why in practice, psychologists and social scientists often fall back on the 19th century conception of utility as a measure of personal wealth or welfare. That gives the MEU Principle considerable predictive power. It also renders the principle obviously false, both as a normative principle about what people should do and as a descriptive principle about people's actual behaviour.

8.6 Further reading

Some good (and rare) discussion of the localism issue can be found in

- Jamie Dreier: "Rational preference: Decision theory as a theory of practical rationality" (1996),
- Lara Buchak: "Redescription", chapter 4 of her *Risk and Rationality* (2013),
- Paul Weirich: "Expected Utility and Risk" (1986).

If you want to know more about Buchak's risk-weighted expected utility theory, have a look at

- Lara Buchak: "Risks and tradeoffs". (2014)

The Jeffrey/Bolker construction is described in

- Richard Jeffrey: *The Logic of Decision*, chapter 9 (1983).

The paper by Harsanyi mentioned in section 8.1 is

- John Harsanyi: "Rule utilitarianism, rights, obligations and the theory of rational behavior" (1980).

Essay Question 8.1

Discuss Buchak's defence of localism in the "Redescription" chapter. What are her main arguments? Can you think of objections or further support?

9 Evidential and Causal Decision Theory

9.1 Evidential decision theory

The traditional method for evaluating an agent's options in a decision situation begins by setting up a decision matrix which identifies the relevant states, acts, and outcomes. The expected utility of each act is then computed as the weighted average of the utility of the possible outcomes, weighted by the probability of the corresponding states.

Finding an adequate decision matrix is not always easy. Among other things, we have to make sure that the propositions we choose as the states are independent of the acts. Exercise 1.8 illustrated why this is needed: a student, wondering whether to study for an exam, drew up the following matrix.

	Will Pass (0.5)	Won't Pass (0.5)
Study	Pass & No Fun (1)	Fail & No Fun (-8)
Don't Study	Pass & Fun (5)	Fail & Fun (-2)

To the student's delight, the expected utility of studying (-3.5) is lower than that of not studying (1.5). The student is wrong, because the states 'Will Pass' and 'Won't Pass' in her matrix are not independent of the acts.

What exactly does independence require? There are several notions of independence.

- Two propositions A and B are **logically independent** if all the combinations $A \wedge B$, $A \wedge \neg B$, $\neg A \wedge B$, and $\neg A \wedge \neg B$ are logically possible.
- A and B are **probabilistically independent** relative to some credence function Cr if $\text{Cr}(B/A) = \text{Cr}(B)$. (See section 2.4.)
- A and B are **causally independent** if, whether or not one of them is true is has no causal influence over whether the other is true.

Exercise 9.1 ⋆

I the student's decision matrix, is it reasonable to assume that the states ('Will Pass', 'Won't Pass') are logically independent of the acts? Is it reasonable to assume that they are causally independent? Is it reasonable to treat them as probabilistically independent?

When we require that the states should be independent of the acts, we don't just mean logical independence. But it is not obvious whether we should require probabilistic independence or causal independence. The question turns out to mark the difference between two fundamentally different approaches to rational choice. If we require probabilistic independence (also known as 'evidential independence'), we get **evidential decision theory** (EDT, for short). If we require causal independence, we get **causal decision theory** (CDT).

Both forms of decision theory say that rational agents maximize expected utility, and they both appear to use the same definition of expected utility: if act A leads to outcomes $O_1, \ldots, O_n$ in states $S_1, \ldots, S_n$ respectively, then

$$EU(A) =_{\text{def}} U(O_1) \cdot \text{Cr}(S_1) + \ldots + U(O_n) \cdot \text{Cr}(S_n).$$

But EDT and CDT disagree on what counts as an adequate state. Each camp accuses the other of making a similar mistake as the student who used 'Will Pass' and 'Won't Pass' as the states. If we require states to be probabilistically independent of the acts, the definition defines **evidential expected utility**; if we require causal independence, it defines **causal expected utility**.

Before we look at examples where the two notions come apart, I want to mention three advantages of the evidential conception.

First, probabilistic independence is much better understood than causal independence. Provided $\text{Cr}(B) > 0$, probabilistic independence between A and B simply means that $\text{Cr}(A) = \text{Cr}(A \wedge B)/\text{Cr}(B)$. By contrast, ever since Hume philosophers have argued that our conception of causality or causal influence is highly problematic. Bertrand Russell, for example, held that "the word 'cause' is so inextricably bound up with misleading associations as to make its complete extrusion from the philosophical vocabulary desirable." All else equal, it would be nice if we could keep causal notions out of our model of rational choice.

A second advantage of EDT is that it allows us to compute expected utilities in a way that is often simpler and more intuitive than the method we've used so far.

Return to the student's matrix. Intuitively, the problem with the matrix is that the 'Will Pass' state is more likely if the student studies than if she doesn't study.

When we evaluate the expected utility of studying, we should therefore give greater weight to worlds in which she passes than when we evaluate the expected utility of not studying.

This suggests that instead of finding a description of the student's decision problem with act-independent states, we might stick with the student's original description, but let the probability of the states vary with the acts. Like so:

	Will Pass	Won't Pass
Study	Pass & No Fun ($U = 1, \text{Cr} = 0.9$)	Fail & No Fun ($U = -8, \text{Cr} = 0.1$)
Don't Study	Pass & Fun ($U = 5, \text{Cr} = 0.2$)	Fail & Fun ($U = -2, \text{Cr} = 0.8$)

'$Cr = 0.9$' in the top left cell indicates that the student is 90% confident that she will pass *if she studies*. By contrast, she is only 20% confident that she will pass *if she doesn't study*, as indicated by '$\text{Cr} = 0.2$' in the bottom left cell. We no longer care about the absolute, unconditional probability of the states. To compute the expected utility of each act we simply multiply the utilities and credences in each cell and add up the products. So the expected utility of studying is $1 \cdot 0.9 + (-8) \cdot 0.1 = 0.1$; for not studying we get $5 \cdot 0.2 + (-2) \cdot 0.8 = -0.6$.

In general, the **new method** for computing expected utilities goes as follows. As before, we set up a decision matrix that distinguishes all relevant acts and outcomes, but we no longer care whether the states are independent of the acts (in any sense). If an act A leads to outcomes $O_1, \ldots, O_n$ in states $S_1, \ldots, S_n$ respectively, then the expected utility of A is computed as

$$EU(A) = U(O_1) \cdot \text{Cr}(S_1/A) + \ldots + U(O_n) \cdot \text{Cr}(S_n/A).$$

Here the unconditional credences $\text{Cr}(S_i)$ in the old method have been replaced by conditional credences $\text{Cr}(S_i/A)$.

Exercise 9.2 ⋆⋆

You have a choice of going to party A or party B. You prefer party A, but you'd rather not go to a party if Bob is there. Bob, however, wants to go where you are, and there's a 50% chance that he will find out where you go. If he does, he will come to the same party, otherwise he will randomly choose one of the two parties. Here is a matrix for your decision problem.

	Bob at A (0.5)	Bob at B (0.5)
Go to A	Some fun (1)	Great fun (5)
Go to B	Moderate fun (3)	No fun (0)

(a) Explain why this is not an adequate matrix for computing expected utilities by the old method.

(b) Use the new method to compute the expected utilities.

We can go one step further. Suppose the outcomes $O_1, \ldots, O_n$ that might result from act A are all distinct, so that A leads to a outcome O_1 only if S_1 obtains, to outcome O_2 only if S_2 obtains, and so on. On the assumption that the agent chooses A, it is then certain that S_1 is true iff O_1 is true, that S_2 is true iff O_2 is true, and so on. It follows that $\text{Cr}(S_1/A) = \text{Cr}(O_1/A)$, $\text{Cr}(S_2/A) = \text{Cr}(O_2/A)$, etc. Substituting these terms in the new formula for $EU(A)$, we get

$$EU(A) = U(O_1) \cdot \text{Cr}(O_1/A) + \ldots + U(O_n) \cdot \text{Cr}(O_n/A).$$

So we can equivalently compute the expected utility of A directly in terms of outcomes, without even mentioning any states.

To derive the equivalence, I have assumed that same outcome never occurs in different states. But the equivalence still holds if we drop that assumption. For example, suppose A leads to O_1 in both S_1 and S_2 (and in no other state). On the assumption that the agent chooses A, it is then certain that O_1 is true iff $S_1 \vee S_2$ is true. Since the states are mutually incompatible, it follows that $\text{Cr}(O_1/A) = \text{Cr}(S_1/A) + \text{Cr}(S_2/A)$. And this means that we can substitute

$$U(O_1) \cdot \text{Cr}(S_1/A) + U(O_1) \cdot \text{Cr}(S_2/A)$$

in the new formula for expected utility by

$$U(O_1) \cdot \text{Cr}(O_1/A).$$

The upshot is that the new method is equivalent to a **state-free method** for computing expected utilities. Here we only need to figure out all the outcomes $O_1, \ldots, O_n$ that a given act might bring about. We then consider how likely each of the *outcomes* is on the supposition that the act is chosen, and take the sum of the products:

$$EU(A) = U(O_1) \cdot \text{Cr}(O_1/A) + \ldots + U(O_n) \cdot \text{Cr}(O_n/A).$$

In practice, this method is often simpler and more intuitive than the old method.

Exercise 9.3 ⋆

You have two options. You can get £10 for sure (utility 1), or flip a fair coin and get £20 on heads (utility 1.8) or £0 on tails (utility 0). The coin will not be flipped if you take the £10. In cases like this, it is hard to find a suitable set of states. Use the state-free method.

I have shown that the new method and the state-free method always yield the same result. I will now show that if we compute the expected utility of an act by one of these methods, what we get is the act's *evidential* expected utility, as defined above. The user-friendliness of the new methods is therefore an argument in favour of EDT.

The proof is easy. The evidential expected utility of an act A is defined as

$$U(O_1) \cdot \mathrm{Cr}(S_1) + \ldots + U(O_n) \cdot \mathrm{Cr}(S_n),$$

where $O_1, \ldots, O_n$ are the outcomes that result from A in states $S_1, \ldots, S_n$ respectively, and the states are probabilistically independent of the acts. By the new method, we would instead compute the expected utility as

$$U(O_1) \cdot \mathrm{Cr}(S_1/A) + \ldots + U(O_n) \cdot \mathrm{Cr}(S_n/A).$$

But since the states are probabilistically independent of the acts, in all these terms, $\mathrm{Cr}(S_i) = \mathrm{Cr}(S_i/A)$. So the two methods here yield the same result. But we know that *any* application of the new method yields the same result as the state-free method. It follows that any application of either the new method or the state-free method yields the same result as the old method as understood by EDT.

A third advantage of EDT is that it provides a powerful argument in support of the MEU Principle, which I mentioned section 8.1. The argument is that the evidential *expected utility* of an act equals the act's *utility*. The principle to maximize evidential expected utility is therefore equivalent to the principle that one should choose acts that are most desirable in light of one's total beliefs and desires. And that sounds very plausible.

Here is another proof that an act's utility equals its evidential expected utility, using the state-free method for computing evidential expected utilities. Suppose A has $O_1, \ldots, O_n$ as (distinct) possible outcomes. Then A is logically equivalent to the disjunction of all conjunctions of A with the outcomes: $(A \wedge O_1) \vee \ldots \vee (A \wedge O_n)$. By Jeffrey's axiom for utility,

$$U(A) = U(A \wedge O_1) \cdot \mathrm{Cr}(A \wedge O_1/A) + \ldots + U(A \wedge O_n) \cdot \mathrm{Cr}(A \wedge O_n/A).$$

Since $\mathrm{Cr}(A \wedge O_i/A) = \mathrm{Cr}(O_i/A)$, this simplifies to

$$U(A) = U(A \wedge O_1) \cdot \mathrm{Cr}(O_1/A) + \ldots + U(A \wedge O_n) \cdot \mathrm{Cr}(O_n/A).$$

Moreover, if the outcomes specify everything that matters to the agent, then $U(A \wedge O_i) = U(O_i)$. So we can simplify once more:

$$U(A) = U(O_1) \cdot \mathrm{Cr}(O_1/A) + \ldots + U(O_n) \cdot \mathrm{Cr}(O_n/A).$$

The right-hand side is the state-free definition of evidential expected utility.

9.2 Newcomb's Problem

In 1960, the physicist William Newcomb invented the following puzzle.

Example 9.1 (Newcomb's Problem)
In front of you are a black box and a transparent box. You can see that transparent box contains £1000. You can't see what's in the black box. You have two options. One is to take (only) the black box and keep whatever is inside. Your second option is to take the black box *and* the transparent box and keep their contents. A demon has tried to predict what you will do. If she predicted that you will take both boxes, then she put nothing in the black box. If she predicted that you will take just the black box, she put £1,000,000 in the box. The demon is very good at predicting this kind of choice. Your options have been offered to many people in the past, and the demon's predictions have almost always been correct.

What should you do, assuming you want to get as much money as possible?

Let's see how EDT and CDT answer the question, starting with CDT. If you only care about how much money you will get, then the following matrix adequately represents your decision problem, according to CDT.

	£1,000,000 in black box	£0 in black box
Take only black box	£1,000,000	£0
Take both boxes	£1,001,000	£1000

Note that the states are causally independent of the acts, as CDT requires: whether you take both boxes or just the black box – in philosophy jargon, whether you

two-box or *one-box* – is certain to have no causal influence over what's in the boxes. This is crucial to understanding Newcomb's Problem. By the time of your choice, the content of the boxes is settled. The demon won't magically change what's in the black box in response to your choice; her only superpower is predicting people's choices.

It is obvious from the decision matrix that taking both boxes maximizes causal expected utility, since it dominates one-boxing: it is better in every state. We don't need to fill in the precise utilities and probabilities.

Turning to EDT, we do need to specify a few more details. Let's say you are 95% confident that there is a million in the black box if you one-box, and 5% confident that there is a million in the black box if you two-box. Let's also assume (for simplicity) that your utility is proportional to the amount of money you will get. Using the new method, the evidential expected utility of the two options then works out as follows ('1B' is one-boxing, '2B' is two-boxing):

$$\begin{aligned} EU(\text{1B}) &= U(£1{,}000{,}000) \cdot \text{Cr}(£1{,}000{,}000/\text{1B}) + U(£0) \cdot \text{Cr}(£0/\text{1B}) \\ &= 1{,}000{,}000 \cdot 0.95 + 0 \cdot 0.05 = 950{,}000. \\ EU(\text{2B}) &= U(£1{,}001{,}000) \cdot \text{Cr}(£1{,}001{,}000/\text{2B}) + U(£1000) \cdot \text{Cr}(£1000/\text{2B}) \\ &= 1{,}001{,}000 \cdot 0.05 + 1000 \cdot 0.95 = 51{,}000. \end{aligned}$$

One-boxing comes out as significantly better than two-boxing.

So CDT says that you should two-box, and EDT says you should one-box. Who has it right? Philosophers have been debating the question for over 50 years, with no consensus in sight.

Some think one-boxing is obviously the right choice. After all, you're almost certain to get more if you one-box than if you two-box. Look at all the people that have been offered the choice in the past. Those who one-boxed almost always walked away with a million, while those who two-boxed walked away with a thousand. Wouldn't you rather be in the first group than in the second? It's your choice!

Others think it equally obvious that you should take both boxes. If you take both boxes, you are guaranteed to get £1000 more than whatever you'd get if you took just the black box. Remember that the content of the boxes is settled. The black box either contains a thousand or a million. And since one-boxing and two-boxing will both give you the black box, it is settled that you will get however much is in that box. The only thing that isn't settled – the only thing over which you have any control – is whether you also get the £1000 from the

transparent box. And if you prefer more money to less money, then clearly (so the argument) you should take the additional £1000.

The argument for two-boxing can be strengthened by the following observation. Imagine you have a friend who helped the demon prepare the boxes. Your friend knows what's in the black box. You've agreed to a secret signal by which she will let you know whether it would be better for you to choose both boxes or just the black box. If you trust your friend, it seems that you should follow her advice. But what will she signal? If the box is empty, she will advise you to take both boxes, so that you get at least the thousand. If the box contains a million, she will also advise you to take both boxes, so that you get £1,001,000 rather than £1,000,000. Either way, she will signal to you that you should take both boxes. But this means you can follow your friend's advice without even looking at her signal. Indeed, you can (and ought to) follow her advice even if she doesn't actually exist.

Why should you follow the advice of your imaginary friend? Think about why we introduced the notion of expected utility in the first place. In chapter 1, we distinguished between what an agent ought to do *in light of all the facts*, and what she ought to do *in light of her beliefs*. In the miner problem (example 1.1), the best choice in light of all the facts is to block whichever shaft the miners are actually in. But since you don't know where the miners are, you don't know what would be the best choice in light of all the facts. You have to go by the limited information you have. The best choice in light of that information is arguably to block neither shaft. But in Newcomb's problem, you actually know what is best in light of all the facts: you know what someone who knows all relevant facts would advise you to do. She would advise you to two-box. (Equivalently, you know what you would decide to do if *you* knew what's in the black box: you would decide to two-box.) Plausibly, if you are certain that some act is best in light of all the facts, then you should choose that act.

Exercise 9.4 ⋆⋆

Show that if you follow EDT, you would not want to know what's in the black box. You'd be willing to pay the demon £500 for not revealing to you the content of the box.

What about the fact that one-boxers are generally richer than two-boxers? Doesn't that show that the one-boxers are doing something right? Not so, say those who advocate two-boxing. The two-boxers who walked away with a mere

thousand were never given a chance to get a million. They were confronted with an empty black box and a transparent box containing £1000; it's hardly their fault that they didn't get a million. On the other hand, all those one-boxers who got a million were effectively given a choice between £1,001,000 and £1,000,000. The fact that they got a million hardly shows that they made the right choice. As an analogy, imagine there are two buttons labelled 'dark' and 'blonde'. If you press the button that matches your hair colour, you get a million if your hair is blonde and a thousand if it is dark. Almost everyone who presses 'blonde' walks away with a million, while almost everyone who presses 'dark' walks away with a thousand. It clearly doesn't follow that everyone should have pressed 'blonde'. Those with dark hair never had a chance to get a million.

9.3 More realistic Newcomb Problems?

Newcomb's Problem is science fiction. Nobody ever faces that situation. Why should we care about the answer?

Philosophers care because the problem brings to light a more general issue: whether the norms of practical rationality must involve causal notions. Those who favour two-boxing in Newcomb's Problem argue that the apparent advantage of EDT, that it does not appeal to causal notions, is actually a flaw. In effect, EDT recommends choosing acts whose choice would be good news. One-boxing in Newcomb's Problem would be good news because it would provide strong evidence that the black box (which you're certain to get) contains a million. By contrast, two-boxing would provide strong evidence that the black box is empty; it would be bad news. But the aim of rational choice, say advocates of CDT, is to *bring about good outcomes*, not to *receive good news*. In Newcomb's Problem, one-boxing is evidence for something good, but it does not contribute in any way to bringing about that good. If the million is in the black box, then it got in there long before you made your choice.

This difference between EDT and CDT can also show up in more realistic scenarios. Some versions of the Prisoner Dilemma (example 1.3) are plausible candidates. Suppose you only care about your own prison term. We can then represent the Prisoner Dilemma by the following matrix, in which the "states" (your partner's options) are causally independent of the acts.

	Partner confesses	Partner silent
Confess	5 years (-5)	0 years (0)
Remain silent	8 years (-8)	1 year (-1)

No matter what your partner does, confessing leads to the better outcome. But now suppose your partner is in certain respects much like you, so that she is likely to arrive at the same decision as you. Concretely, suppose you are 80% confident that your partner will choose whatever you will choose, so that Cr(she confesses/you confess) = Cr(she is silent/you are silent) = 0.8. As you can check, EDT then recommends remaining silent. Friends of CDT think that this is wrong. Under the given assumptions, remaining silent is good news, as it indicates that your partner will also remain silent – and note how much better the right-hand column is than the left-hand column. But that is no reason for you to remain silent.

Exercise 9.5 ⋆

Compute the evidential expected utility of confessing and remaining silent.

Another potential example are so-called **Medical Newcomb problems**. In the 1950s, it became widely known that the cancer rate is a lot higher among smokers than among non-smokers. Fearing that a causal link between smoking and cancer would hurt their profits, tobacco companies promoted an alternative explanation for the finding. The correlation between smoking and cancer, they suggested, is due to a common cause: a genetic disposition that causes both a desire to smoke and cancer. The cancer, on that explanation, isn't caused by smoking, but directly by the genetic factors that happen to also cause smoking.

Why would the tobacco industry be interested in promoting this hypothesis? Because they assumed that if people believed that smoking does not actually increase the risk of cancer, but merely indicates a genetic predisposition for cancer, then people would keep smoking. According to EDT, however, it seems that people should give up smoking either way, for on either hypothesis smoking is bad news.

Let's work through an example. Suppose you assign some (sub)value to smoking, but greater (sub)value to not having cancer, so that your utilities for the

possible combinations of smoking and getting cancer are as follows:

$$\begin{aligned} U(\text{smoking} \wedge \neg\text{cancer}) &= 1 \\ U(\neg\text{smoking} \wedge \neg\text{cancer}) &= 0 \\ U(\text{smoking} \wedge \text{cancer}) &= -9 \\ U(\neg\text{smoking} \wedge \text{cancer}) &= -10 \end{aligned}$$

Suppose you are convinced by the tobacco industry's explanation: you are sure that smoking does not cause cancer. But you think smoking is evidence for the cancer gene. So Cr(cancer/smoking) > Cr(cancer/¬smoking). Let's say Cr(cancer/smoking) = 0.8 and Cr(cancer/¬smoking) = 0.2. It follows that the evidential expected utility of smoking is $-9 \cdot 0.8 + 1 \cdot 0.2 = -7$, while the evidential expected utility of not smoking is $-10 \cdot 0.2 + 0 \cdot 0.2 = -2$. According to EDT, then, you should stop smoking even if you buy the tobacco industry's explanation. Indeed, it should make no difference to you whether smoking causes cancer or merely indicates a predisposition for cancer.

That is not what the tobacco industry expected. And it does seem odd. In the example, you are sure that smoking will not bring about anything bad. On the contrary, it is guaranteed to make things better. At the same time, it would be evidence that you have the cancer gene. By not smoking, you can suppress this piece of evidence, but you can't affect the likelihood of getting cancer. If what you really care about is whether or not you get cancer, rather than whether or not you *know* that you get cancer, what's the point of making your life worse by suppressing the evidence?

Friends of EDT have a response to this kind of example. If the case is to be realistic, they have argued, smoking actually won't be evidence for cancer: Cr(cancer/smoking) won't be greater than Cr(cancer/¬smoking). For we've assumed that the gene causes smoking by causing a desire to smoke. But suppose you feel a strong desire to smoke. That desire provides evidence that you have the gene. Acting on the desire would provide no further evidence. Similarly if you don't feel a desire to smoke: not feeling the desire is evidence that you don't have the gene, and neither smoking nor not smoking then provides any further evidence. So once you've taken into account the information you get from the presence or absence of the desire, Cr(cancer/smoking) = Cr(cancer/¬smoking), and then EDT recommends smoking.

This response has come to be known as the "tickle defence" of EDT, because it

assumes that the cancer gene would cause a noticeable "tickle" whose presence or absence provides all the relevant evidence.

Exercise 9.6 ⋆

You wonder whether to vote in a large election between two candidates A and B. You assign (sub)value 100 to A winning and 0 to B winning. Voting would add a (sub)value of -1, since it would cause you some inconvenience. Your credence that your vote will make a difference is 0.001. You figure out that not voting maximizes expected utility. But then you realize that other potential voters are likely to go through the same thought process as you. You estimate that around 1% supporters of A might go through the same process of deliberation as you, and will reach the same conclusion that you will reach. Does that change the causal expected utility of voting? Does it change the evidential expected utility? (Explain briefly, without computing anything.)

9.4 Causal decision theories

Those who are convinced by the case against EDT believe that some causal notion must figure in an adequate theory of rational choice: rational agents maximize causal expected utility.

One way to define causal utility is the classical definition in terms of states, acts, and outcomes, where the states are required to be causally independent of the acts. But we can also construct a version of CDT that looks more like EDT and shares at least some of the attractive features of EDT. The key to this construction is a point I mentioned in section 2.4: that there are two ways of supposing a proposition.

What would have happened if the Nazi program to build nuclear weapons had succeeded in 1944? When we contemplate this question, we consider possible histories that are like the history of our world until 1944 but then depart in some minimal way to allow the Nazi's nuclear weapon program to succeed. (The departure should be "minimal" because we're not interested in worlds where, say, the Nazis learned how to build nuclear weapons from gigantic aliens who invaded the Earth in 1944, destroying the continent of America as they landed.) After that departure, the alternative histories should continue to develop in accordance with the general laws of our world. Since Hitler's character is the same in the

alternative worlds and in the actual world, it seems likely that Hitler would have used the nuclear weapons, possibly leading to an Axis victory in World War II.

This is an example of **subjunctive supposition**. In general, in subjunctively supposing an event, we consider what a world would be like that closely resembles the actual world up to the relevant time, then departs minimally to allow for the event, and afterwards develops in accordance with the general laws of the actual world. Subjunctive supposition is a causal kind of supposition. When we subjunctively suppose that the Nazis had nuclear weapons, we consider what this event *would have brought about*.

By contrast, when we **indicatively suppose** that the Nazis had nuclear weapons, we hypothetically add the supposed proposition to our beliefs and revise the other beliefs in a minimal way to restore consistency. We do not suspend our belief that the Nazis lost the war, that they did not use nuclear weapons, etc. On the indicative supposition that the Nazis had nuclear weapons in 1944, we conclude that something prevented the use of the weapons, an act of sabotage perhaps.

In a probabilistic framework, $\mathrm{Cr}(B/A)$ is an agent's credence in B on the indicative supposition that A; if $\mathrm{Cr}(A) > 0$ it equals $\mathrm{Cr}(B \wedge A)/\mathrm{Cr}(A)$. Let '$\mathrm{Cr}(B//A)$' (with two dashes) denote an agent's credence in B on the *subjunctive* supposition that A. There is no simple analysis of $\mathrm{Cr}(B//A)$ in terms of the agent's credence in B, A, or logical combinations of these. Whether A would bring about B is generally not a matter of logic, but depends on the laws of nature and various particular facts besides A and B.

Some have suggested that the probability of B on the subjunctive supposition that A equals the probability of the corresponding *subjunctive conditional*: the sentence 'if A were the case then B would be the case'. So the proposal is that $\mathrm{Cr}(B//A) = \mathrm{Cr}(A \mathrel{\square\!\!\rightarrow} B)$, where '$A \mathrel{\square\!\!\rightarrow} B$' abbreviates 'if A were the case then B would be the case'. Whether this is a step forward depends on what more we can say about the proposition $A \mathrel{\square\!\!\rightarrow} B$. We won't pursue the question any further.

Now return to the new method for computing expected utilities from section 9.1. The idea was to use conditional probabilities instead of unconditional probabilities, which allowed us to drop the requirement that states and acts are independent:

$$EU(A) = U(O_1) \cdot \mathrm{Cr}(S_1/A) + \ldots + U(O_n) \cdot \mathrm{Cr}(S_n/A).$$

Here the conditional probabilities are indicative. But we can just as well use subjunctive conditional probabilities, considering what the relevant act A would

be likely to bring about:

$$EU(A) = U(O_1) \cdot \mathrm{Cr}(S_1 // A) + \ldots + U(O_n) \cdot \mathrm{Cr}(S_n // A).$$

As before, one can show (under plausible assumptions) that this method for computing expected utilities yields the same result as our original definition of causal expected utilities. It is also equivalent to a state-free method:

$$EU(A) = U(O_1) \cdot \mathrm{Cr}(O_1 // A) + \ldots + U(O_n) \cdot \mathrm{Cr}(O_n // A).$$

To get a feeling for how this works, let's first apply it to a simple case inspired by Newcomb's problem. Depending on the outcome of a coin toss, a box has been filled with either £1,000,000 or £0. You can take the box or leave it. How much would you get if you were to take the box? It depends on what's inside. If the box contains £1,000,000, then you would get £1,000,000 on the (subjunctive) supposition that you take the box. If the box contains £0, you would get £0 if you were to take the box. Both possibilities have equal probability, so

$$\mathrm{Cr}(\text{£1,000,000} // \text{Take box}) = 0.5$$
$$\mathrm{Cr}(\text{£0} // \text{Take box}) = 0.5.$$

In general, if a box contains a certain amount of money, and you have the option of taking the box, and you are certain that taking the box would not alter what's inside the box, then on the subjunctive supposition that you take the box, you are certain to get however much is in the box. Any uncertainty about how much you would get boils down to uncertainty about how much is in the box.

Let x be your credence that the black box in Newcomb's Problem contains a million. Accordingly,

$$\mathrm{Cr}(\text{£1,000,000} // 1B) = x;$$
$$\mathrm{Cr}(\text{£0} // 1B) = 1 - x;$$
$$\mathrm{Cr}(\text{£1,001,000} // 2B) = x;$$
$$\mathrm{Cr}(\text{£1000} // 2B) = 1 - x.$$

Using the new method for computing causal expected utilities, it follows that

$$EU(1B) = 1,000,000 \cdot x + 0 \cdot (1 - x);$$
$$EU(2B) = 1,001,000 \cdot x + 1000 \cdot (1 - x).$$

No matter what x is, taking both boxes maximizes (causal) expected utility.

Exercise 9.7 ⋆⋆⋆

Consider the third argument in favour of EDT from section 9.1: that an act's evidential expected utility equals the act's utility. Can we adapt this line of argument to CDT? How would we have to change the theory of utility from section 5.3?

9.5 Unstable decision problems

A curious phenomenon that can arise in CDT is that the choiceworthiness of an option changes during deliberation.

Example 9.2

There are three boxes: one red, one green, one transparent. You can choose exactly one of them. The transparent box contains £100. A demon with great predictive powers has anticipated your choice. If she predicted that you would take the red box, she put £200 in the red box and £250 in the green box. If she predicted that you would take the green box, she put £0 in the green box and £100 in the red box. If she predicted that you would take the transparent box, she put £90 in both the red and the green box.

Here is a matrix for the example. 'R', 'G', 'T' are the three options (red, green, transparent).

	Predicted R	Predicted G	Predicted T
R	£90	£130	£100
G	£110	£80	£90
T	£100	£100	£100

Let's say you initially assign equal credence to the three predictions, and your utility for money is proportional to the amount of money. It is easy to see that R then maximizes (causal) expected utility. So you should be tempted to choose the red box. But if you are tempted to choose the red box, then it is no longer rational to treat all three predictions as equally likely: you should become more confident that the demon predicted R. But if you are sufficiently confident that the demon predicted R, then R no longer maximizes expected utility! Rather, you should then go for G. But if you're tempted to go for G, then again you should adjust your credences, and so on.

Exercise 9.8 ⋆⋆

Can you see where this process of deliberation will end? (Explain briefly.)

It can even happen that whatever option you currently favour makes an alternative option look more appealing, so that it becomes impossible to reach a decision.

Example 9.3 (Death in Damascus)

At a market in Damascus, a man meets Death. Death looks surprised; "I am coming for you tomorrow", he says. Terrified, the man buys a horse and rides all through the night to Aleppo. The next day, Death knocks on the door of the room where the man is hiding. "I was surprised to see you in Damascus", Death explains, "for I knew I had an appointment with you here today."

Suppose you're the man in the story, having just met Death in Damascus. Death has predicted where you are going to be tomorrow. Let's assume the prediction is settled, and not affected by what you decide to do. But Death is a very good predictor. So if you go to Aleppo, you can be confident that Death will wait for you there, while if you stay in Damascus, you can be confident that Death will be in Damascus. The more you are inclined towards one option, the more attractive the other option becomes.

If we interpret the MEU Principle causally, then our model of rationality seems to rule out both options in *Death in Damascus*: you can't rationally choose to go to Aleppo, for then you should be confident that Death will wait in Aleppo, in which case staying in Damascus maximizes expected utility; for parallel reasons, you also can't rationally choose to stay in Damascus. But you only have these two options! How can both of them be wrong?

Exercise 9.9 ⋆⋆

Let's modify the *Death in Damascus* scenario by assuming that staying in Damascus would have some benefits, independently of whether Death will find you. So the matrix might look like this:

	Death in Aleppo	Death in Damascus
Go to Aleppo	death (-100)	life (5)
Stay in Damascus	life & benefits (10)	death & benefits (-95)

Now what does CDT say you should do? What about EDT?

9.6 Further reading

The dispute over Newcomb's Problem has become quite sophisticated, as you can see from the following articles (the first of which defends one-boxing, the second two-boxing):

- Arif Ahmed: "Causation and Decision" (2010)
- Jack Spencer and Ian Wells: "Why Take Both Boxes?" (2017)

The classical treatment of EDT is Richard Jeffrey's *Logic of Decision* (1965/1983). A classical exposition of CDT is

- David Lewis: "Causal Decision Theory" (1981).

Lewis argues that various methods for computing causal expected utility are plausibly equivalent.

The model of deliberation outlined in section 9.5 is due to Brian Skyrms. A good introduction can be found in

- Frank Arntzenius: "No Regrets, or: Edith Piaf Revamps Decision Theory" (2008).

Essay Question 9.1

Should a rational agent take both boxes in Newcomb's Problem or just the black box? Can you think of an argument for either side not mentioned in the text?

10 Game Theory

10.1 Games

Game theory studies situations in which an agent faces one or more choices whose combined outcome depends on the choices of other agents. Such situations are called **games**, and the agents **players**. The Prisoner Dilemma (example 1.3) is a game in this sense, because the outcome of your choice (confessing or remaining silent) depends on what your partner decides to do.

Whenever an agent faces a choice in a game, the MEU Principle tells us that she ought to choose whichever option maximizes expected utility. We don't need a new decision theory for games. Nonetheless, there are reasons for studying the special case where the states in a decision problem are other people's (real or potential) actions.

One reason is that game theory can shed light on important political and social issues. The Prisoner Dilemma, for instance, illustrates a common type of situation in which everyone pursuing their own interest leads to a worse result for everyone than what could otherwise have been achieved. Any such situation is nowadays called a Prisoner Dilemma, even if no prisoners are involved. For example, if you're a professional athlete, you have an incentive to use steroids, no matter whether your competitors do the same. Without strict controls, the outcome is that everyone uses steroids, even though everyone would prefer that no-one uses steroids. The athletes face a Prisoner Dilemma. For another example, if you're a fisherman, you have an incentive to catch as many fish as you can, even though everyone would be better off if everyone restrained themselves to sustainable quotas.

Thomas Hobbes (in effect) argued that the pervasiveness of Prisoner Dilemmas justifies the subordination of people under a state. It is in everyone's interest to impose a system of control and punishment that ensures the best outcome in what would otherwise be a Prisoner Dilemma.

Another reason to study games is that a new set of conceptual tools and

techniques become available if the outcomes depend on other people's choices. For example, game theorists typically don't specify the probability of the states – that is, of the other players' actions. Instead, they specify the utility of the outcomes for all players. Under certain assumptions, this is enough to figure out what each player will do.

Here is how game theorists would write the matrix for the original Prisoner Dilemma, assuming you and your partner only care about your own prison terms:

	Confess	Silent
Confess	-5,-5	0,-8
Silent	-8,0	-1,-1

The rows are the acts available to you, as before; the columns are the acts available to your partner. The numbers in the cells represent the utility of the relevant outcome for you and your partner. We usually don't describe the outcome itself anymore (for lack of space). In a two-player matrix, the first number in each cell is always the utility for the row player (whom we'll call 'Row'); the second is the utility for the column player ('Column').

In the above matrix, it is clear that each player will confess, because confessing dominates remaining silent: it is better no matter what the other player does.

Now consider the following matrix, for a different kind of game.

	C_1	C_2
R_1	2,2	1,3
R_2	1,1	2,2

Row no longer has a dominant option. What she should do depends on what she thinks Column will do. If Column chooses C_1, then Row should play R_1; if Column chooses C_2, then Row should play R_2. Can we nonetheless say what Row will do without specifying her beliefs?

Look at the game from Column's perspective. No matter what Row does, Column is better off choosing C_2. In other words, C_2 dominates C_1. So if Row knows the utility Column assigns to the outcomes, she can figure out that Column will choose C_2. And so Row should choose R_2. The "solution" of the game is therefore the pair of options (R_2, C_2) – meaning that if Column and Row are rational players knowing about each other's utilities, then Row will choose R_2 and Column C_2.

Here is another, more complex example.

	C_1	C_2	C_3
R_1	0,1	2,2	3,1
R_2	2,2	1,3	2,2
R_3	1,1	0,2	0,3

From Row's perspective, R_1 is the best choice if Column plays C_2 or C_3, and R_2 is the best choice if Column goes for C_1. For Column, C_2 is the best choice in case of R_1 or R_2, and C_3 is best in case of R_3. But Column can hardly expect Row to choose R_3, since R_3 is dominated by R_2. So Column can figure out that Row will play either R_1 or R_2, which means that Column will play C_2. And since Row can figure out that Column will play C_2, Row will play R_1. The solution is (R_1, C_2).

Note that to reach this conclusion it is not enough to assume that both players know each other's utilities. For one thing, the players must also know that none of them will choose a dominated act. Moreover, to figure out that Column will play C_2, Row needs to know that Column knows her (Row's) utilities, and she needs to know that Column knows that she (Row) won't choose a dominated option.

In general, game theorists usually assume that

(1) all players know the options and utilities of all other players;
(2) all players know that all other players are rational;
(3) all players know that (1)–(3) are satisfied.

By applying to itself, the last clause ensures that (1) and (2) hold with arbitrarily many iterations of 'all players know that' stacked in front. If something is in this way known by everyone, and known by everyone to be known by everyone, and so on, then it is said to be **common knowledge**. So (1)–(3) say that the options and utilities of all players, as well as all players' rationality, are common knowledge. This is obviously an idealization. We'll see below that it gives rise to some interesting puzzles.

Exercise 10.1 ⋆⋆

Under the assumptions (1)–(3), what will Row and Column do in the following games?

a.

	C_1	C_2
R_1	1,0	1,2
R_2	0,3	0,1

b.

	C_1	C_2	C_3
R_1	1,0	1,2	0,1
R_2	0,3	0,1	2,0

c.

	C_1	C_2	C_3
R_1	0,1	2,0	2,4
R_2	4,3	1,4	2,5
R_3	2,4	3,6	3,1

10.2 Nash equilibria

Have a look at this game.

	C_1	C_2	C_3
R_1	4,2	2,3	3,1
R_2	3,1	3,2	4,1
R_3	4,2	1,1	0,3

No option for either player is dominated by any other. Nonetheless, game theory says that we can figure out what each player will choose.

Let's start with some trial and error. Take (R_1, C_1). Could this be the outcome that is reached whenever the game is played by two players under the idealizing assumptions (1)–(3)? No. Otherwise Column would know that Row is going to play R_1. And then Column is better off playing C_2. The opposite happens with (R_1, C_2): if Row knew that Column plays C_2, she would be better off playing R_2. The same reasoning disqualifies all other combinations except (R_2, C_2) – the middle cell. If Row knows that Column is going to play C_2, she can do no better than play R_2. Likewise for Column: if Column knows that Row is going to play R_2, she can do no better than play C_2.

A combination of options that is "stable" in this way is called a **Nash equilibrium** (after the economist John Nash). In general, a Nash equilibrium is a combination of acts, one for each player, such that no player could get greater utility by deviating from her part of the equilibrium given that the other players stick to their part.

Here is an algorithm for finding Nash equilibria. Start from the perspective of the row player. For each act of the column player, underline the best outcome(s) Row can achieve if Column chooses this act. In the example above, you would underline the 4s in the first column, the 3 in the middle cell, and the 4 in the third column. Then do the same with the column player: for each act of Row, underline the best possible outcome(s) for Column. The result looks like this.

	C_1	C_2	C_3
R_1	$\underline{4}$,2	2,$\underline{3}$	3,1
R_2	3,1	$\underline{3}$,$\underline{2}$	$\underline{4}$,1
R_3	$\underline{4}$,2	1,1	0,$\underline{3}$

Any cell in which both numbers are underlined identifies a Nash equilibrium.

According to standard game theory, if a game has a unique Nash equilibrium, and assumptions (1)–(3) hold, then the players will play their part of the equilibrium.

This is not as obvious as it may perhaps appear. Consider the next game.

	C_1	C_2	C_3
R_1	$\underline{2}$,-2	-1,$\underline{1}$	$\underline{1}$,-1
R_2	0,0	$\underline{0}$,0	-2,$\underline{2}$
R_3	0,$\underline{0}$	$\underline{0}$,$\underline{0}$	$\underline{1}$,-1

Here (R_3, C_2) is the unique Nash equilibrium. According to game theory, if you're Row, you can therefore be sure that Column will play C_2. But if you're sure that Column will play C_2, then R_2 and R_3 have equal expected utility! So it is not obvious that you have to play R_3.

To be sure, if you played R_2 and Column could predict your choice, then Column would play C_3, leaving you worse off. But we're not assuming that Column can predict your choice. All we're assuming is (1)–(3). Still, there is an argument in favour of (R_3, C_2). Suppose for reductio that under assumptions (1)–(3) you could just as well play R_2 or R_3. Then Column couldn't be sure which of these you choose; she would have to give roughly equal credence to R_2 and R_3. But then it is best for her to choose C_3. Anticipating this, you would then have to choose R_3. This contradicts our assumption that you could just as well play R_2 or R_3.

Exercise 10.2 ⋆

Identify the Nash equilibria in the following games.

a.

	C_1	C_2
R_1	3,4	4,3
R_2	1,3	5,2
R_3	2,0	1,5

b.

	C_1	C_2	C_3
R_1	1,0	1,2	0,1
R_2	0,3	0,1	2,0

c.

	C_1	C_2	C_3
R_1	0,1	2,0	2,4
R_2	4,3	1,4	2,5
R_3	2,4	3,6	3,1

Exercise 10.3 ⋆⋆

Whenever the method from section 10.1, which is called **elimination of dominated strategies**, identifies a combination of acts as a game's solution, then that combination of acts is a Nash equilibrium. Can you explain why?

10.3 Zero-sum games

In some games, the players' preferences are exactly opposed: if Row prefers a given outcome to another by a certain amount, then Column prefers the second outcome to the first by the same amount. The utilities in every cell therefore sum to the same amount. Since utility scales don't have a fixed zero, game theorists usually re-scale the utilities so that the sum is zero. For that reason, games in which the agents' preferences are exactly opposed are called **zero-sum games**. Here is an example.

	C_1	C_2	C_3
R_1	1,<u>-1</u>	<u>3</u>,-3	<u>1</u>,<u>-1</u>
R_2	<u>2</u>,-2	-2,<u>2</u>	-1,1

(R_1, C_3) is the unique Nash equilibrium. It is also a combination of the players' maximin strategies. Recall that according to the maximin rule, you should choose whichever option has the best worst-case outcome. In the example, the worst-case outcome of R_1 (for Row) is utility 1; for R_2 it is -2. So maximin says that Row should choose R_1. Similarly, it says that Column should choose C_3. And that's just our Nash Equilibrium.

In section 1.4, I argued that maximin is an indefensible decision rule. But it makes sense in the context of zero-sum games, under assumptions (1)–(3). The reason is that whatever you find to be the best option, your opponent can reproduce your reasoning and make sure that she gets the most out of your choice. Hence you can be confident that whatever row you choose, you'll end up in the

worst cell on that row. So you better choose the row where the worst cell is the least bad.

Many games have more than one Nash equilibrium. As we will see, it is then hard to predict what the agents will do without looking at their beliefs, and we can't always assume they will reach a Nash equilibrium. In zero-sum games, however, things are easier. Consider the following example.

	C_1	C_2	C_3
R_1	2,-2	1,-1	1,-1
R_2	3,-3	1,-1	1,-1
R_3	0,0	-1,1	-2,2

The game has four Nash equilibria. So what will the players do? Should Row play R_1 or R_2? Should Column play C_2 or C_3? The answer is that it doesn't matter: they can arbitrarily choose among these options. Whatever they choose, they are guaranteed to end up at an equilibrium, and all the equilibria have the same utility.

Exercise 10.4 ⋆⋆⋆

Prove that this holds for all two-player zero-sum games: if (R_i, C_j) and (R_n, C_m) are Nash equilibria, then so are (R_i, C_m) and (R_n, C_j); moreover, all Nash equilibria have the same utility.

Some games have no Nash equilibrium at all. Here is a matrix for Rock–Paper–Scissors, assuming both players only care about whether they will win or lose.

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

There is no equilibrium. So what should rational players do?

Game theory says that they should randomly choose one of the three options, giving each option the same chance of being chosen. Interpreted literally, this seems to imply that the agents actually have further options besides Rock, Paper, and Scissors: to flip a coin or toss a die and let the result decide what to do. Such a randomized choice is called a **mixed strategy**. The strategy of playing Rock, Paper, or Scissors each with probability 1/3 would be written '[1/3 Rock, 1/3 Paper, 1/3 Scissors]'.

Suppose two players both play [1/3 Rock, 1/3 Paper, 1/3 Scissors]. Then neither could do better by playing anything else (including other mixed strategies). Hence the combination of the two mixed strategies is a Nash Equilibrium. It is the only Nash Equilibrium in Rock–Paper–Scissors.

It can be shown that every game has at least one Nash Equilibrium if mixed strategies are included. (This was shown by John Nash.) The proof obviously assumes that randomization introduces no additional costs or benefits. If you like being in control and therefore prefer losing in Rock–Paper–Scissors to randomizing, then the game has no Nash Equilibrium, not even among mixed strategies.

Exercise 10.5 ⋆⋆

Suppose your opponent plays [1/3 Rock, 1/3 Paper, 1/3 Scissors]. What is the expected utility of playing Rock? How about Paper and Scissors? What is the expected utility of playing [1/3 Rock, 1/3 Paper, 1/3 Scissors]?

10.4 Harder games

Most games in real life are not zero-sum games. The following example illustrates the class of **coordination problems** in which several agents would like to coordinate their actions.

Example 10.1

You and your friend Bob want to meet up, but neither of you knows to which party the other will go. Party A is better than party B, but you will both go home if you don't find each other.

	Party A	Party B
Party A	3,3	0,0
Party B	0,0	2,2

In the example, there are two Nash equilibria (without randomization): both going to party A, and both going to party B. But we can't assume that whenever rational agents play the game, then they will end up in one of these equilibria. If you suspect that Bob will go to party B, and Bob suspects you will go to party A, then you'll go to B and Bob to A.

But could this actually happen, under assumptions (1)–(3)? As you may check, going to party B maximizes expected utility if and only if your credence that Bob

goes to B is at least 0.6. But could you be at least 60% confident that Bob will go to B, given what you know about Bob's utilities? Well, Bob will go to B provided that *he* is at least 60% confident that *you* will go to B. So to be at least 60% confident that Bob will go to B, you only need to be at least 60% confident that Bob is at least 60% confident that you will go to B. Of course, Bob can figure out that you will go to B only if you are at least 60% confident that he will go to B. So to be at least 60% confident that Bob will go to B, you need to be at least 60% confident that Bob is at least 60% confident that you are at least 60% confident that Bob will go to B. And so on. There is nothing incoherent about this state of mind. Nonetheless, we may wonder how you could have arrived at it. How could you have rationally arrived at a 60% confidence that Bob is at least 60% confident that you are at least 60% confident that ... and so on and on forever?

Our assumptions (1)–(3) here give rise to a general epistemological puzzle. If you have no further relevant evidence, how confident should you be that Bob goes to B? You might think your degree of belief should be $1/2$, by the Principle of Indifference. But then you should assume that Bob's degree of belief in *you* going to B is also $1/2$. And that would imply that Bob goes to A. So it can't be right that you should give equal credence to the two possibilities.

In real coordination problems, the players often do have further information. For example, when you're driving on a road, you're typically in a coordination game with drivers going in the opposite direction: you prefer to drive on the left if and only if the others drive on the left; the others prefer to drive on the left if and only if you drive on the left. The existence of a law to drive on the left gives you reason to think that the others will drive on the left. But even without a law, the mere observation that people generally drive on the left would give you reason to think that that's what they will continue to do.

A different kind of coordination is called for in the following game.

Example 10.2 (Chicken)

For fun, you and your friend Bob drive towards each other at high speed. If one of you swerves and the other doesn't, the one who swerves loses. If neither swerves, you both die.

	Swerve	Straight
Swerve	0,0	-1, 1
Straight	1,-1	-10,-10

Games like chicken are sometimes called **anti-coordination games**, because each player would prefer the other one to yield without yielding herself. There are two Nash Equilibria in Chicken that don't involve randomization, (Swerve, Straight) and (Straight, Swerve). As before, either choice is rationally defensible, given suitable beliefs about the opponent, and as before there is an epistemological puzzle about how any of these beliefs could come about.

An interesting feature of many anti-coordination games is that they seem to favour agents who do not maximize expected utility. Suppose Bob is insane and will go straight no matter what, despite the large cost of dying if you both go straight. And suppose you know about Bob's insanity. Then you, as an expected utility maximizer, will have to swerve. And so Bob will win.

There are stories that during the cold war, the CIA leaked false information to the Russians that the US President was an alcoholic, while the KGB falsified medical reports suggesting that Brezhnev was senile. Both sides tried to gain a strategic advantage over the other by indicating that they would irrationally retaliate against a nuclear strike even if they had nothing to gain any more.

Exercise 10.6 ⋆

What should you do in Chicken if you give equal credence to the hypotheses that Bob will swerve and that he will go straight?

Exercise 10.7 ⋆⋆⋆

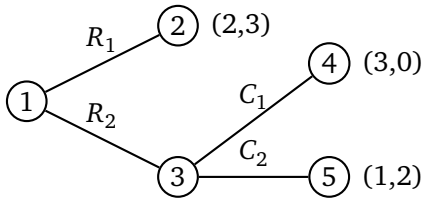
A third Nash equilibrium in Chicken involves randomization. Can you find it? What is the expected utility for both players if they play that mixed strategy?

10.5 Games with several moves

So far we've looked at games in which each player makes just one move and neither player knows about the other's move ahead of their choice. Game theory also studies games in which these assumptions are relaxed. Let's have a quick look at games with several moves, assuming players always know what was played before.

The standard representation of such games are tree-like diagrams known as **extensive form representations**. In the game represented by the diagram below, Row first has a choice between R_1 and R_2. If she chooses R_1 the game ends at node 2 with an outcome that has utility 2 for Row and 3 for Column. If Row

chooses R_2, Column gets a choice between C_1 and C_2. If Column chooses C_1, Row gets utility 3 and Column 0; if Column chooses C_2, Row gets 1 and Column 2.



From a decision theoretic perspective, a game with several moves is a series of (potential) decision problems, one for each occasion where a player faces a choice. The states in these decision problems specify the future decisions of all players. As before, under the assumption that the utilities and rationality of players are common knowledge, we can often predict what will happen without specifying the players' degrees of belief.

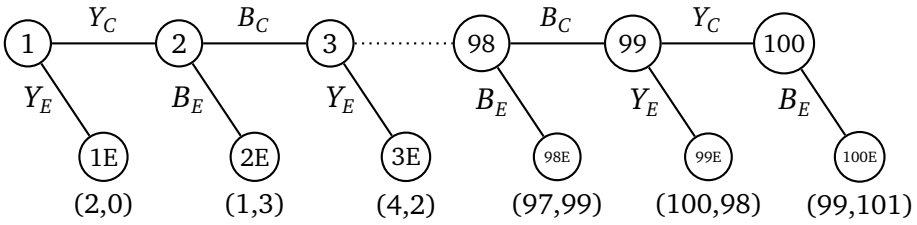
Consider the node labelled '3' in the above tree. Here, Column faces a choice between outcome 4 and outcome 5. That choice involves no relevant uncertainty, and Column prefers outcome 5. So Column can be expected to play C_2. Anticipating this, Row can figure out that if she plays R_2, then outcome 5 will come about, which has utility 1 for Row. By comparison, if she plays R_1, she guarantees outcome 2, which has utility 2. So Row will play R_1.

This process for solving games with multiple moves is called **backward induction**. It can lead to very strange results.

Example 10.3 (Centipede)

You and Bob are playing a game. Initially, there's a pot containing £2. In round 1, you begin by deciding whether to continue or end the game. If you end the game, you get the £2 and Bob gets £0. If you continue, the money in the pot increases by £2 and Bob decides whether to continue or end. If he ends the game here (in round 2), he gets £3 and you get £1. If he continues, the money in the pot increases by £2 and it's your turn again. If you end the game (in round 3), you get £4 and Bob gets £2. And so on. In each round, the money in the pot increases by £2 and whoever ends the game gets £2 more than the other player. In round 100, Bob no longer has an option to continue.

Suppose you and Bob don't care about each other; each of you only wants to get as much money as possible. Here is a partial diagram of the resulting game of Centipede.



We can use backward induction to figure out how you should play. The latest point at which a player has a real choice is round 99. Here, you can either end the game (Y_E) and get £100 or continue (Y_C) and get £99. So you should end the game. Anticipating this, what should Bob do in round 98? If he ends the game (B_E), he'll get £99; if he continues (B_C), he'll get £98. So he should end the game. Anticipating this, you should end the game in round 97, to ensure that you'll get £98 rather than £97. And so on, all the way back to round 1. At each point, backward induction tells us the game should be ended. Thus knowing at round 1 that Bob will end the game in round 2, you should end the game right away. So the "solution" is that you will get £2 and Bob gets £0.

When actual people play the Centipede game, almost no-one ends the game right away. Is this a sign of either altruism or irrationality? Not necessarily.

Let's look at your choice in round 1 from a decision theoretic perspective. It is clear what happens if you end the game: you'll get £2. But what would happen if you decide to continue? The argument from backward induction assumes that Bob would end the game. And if you could be certain that Bob would do that, then you should indeed end the game in round 1. But why should Bob end the game? Because, so the argument, he can be certain that you would end the game in round 3. But the argument for ending in round 3 is exactly parallel to the argument for ending in round 1. Yet if Bob faces a choice in round 2, then he has just seen that you *didn't* end the game in round 1. So he arguably can't be sure you would end it in round 3. On the contrary, he should be somewhat confident that you will continue in round 3. And then Bob should continue in round 2. It follows that continuing in round 1 maximizes expected utility, as it is likely to get you at least to round 3!

Exercise 10.8 ⋆⋆

Change the Centipede game so that there's no fixed end point. Rather, each time a player chooses to continue, the game ends with a probability of 1%.

Does this change anything? How should you play?

Exercise 10.9 ⋆⋆

Suppose you repeatedly face the Prisoner Dilemma with the same partner, for an unknown number of rounds. You only care about your own prison terms. You expect that your partner will remain silent in the first round and from then on imitate whatever you did in the previous round. In this case, you should arguably remain silent. Does that mean you should choose an act that doesn't maximize expected utility?

10.6 Evolutionary game theory

One of the most successful applications of game theory lies (somewhat surprisingly) in the study of biological and cultural evolution. Consider the following game.

Example 10.4 (The Stag Hunt)

Two players independently decide whether to hunt stag or rabbit. Hunting stag requires cooperation, so if only one of the players decides to hunt stag, she will get nothing. The utilities are as follows.

	Stag	Rabbit
Stag	5,5	0,1
Rabbit	1,0	1,1

In the evolutionary interpretation, the utilities represent the *relative fitness* that results from a combination of choices, measured in terms of average number of surviving offspring. Let's assume that each strategy is played by a certain fraction of individuals in a population. Individuals who achieve an outcome with greater utility will, by definition, have more offspring on average, so their proportion in the population will increase.

Suppose initially $1/4$ of the individuals in the population goes for stags and $3/4$ for rabbits. Assuming that encounters between individuals are completely random, this means that any given individual has a $1/4$ chance of playing with someone hunting stag, and a $3/4$ chance of playing with someone hunting rabbit. The average utility of hunting stag is therefore $1/4 \cdot 5 + 3/4 \cdot 0 = 1.25$; for hunting

rabbit the utility is of course 1. Individuals going for stag therefore have greater average fitness. Their fraction in the population increases. As a consequence, it becomes even more advantageous to go for stag. Eventually, everyone will hunt stag.

By contrast, suppose initially only $1/10$ of the population goes for stags. Then hunting stag has an average utility of 0.5, which is less than the utility of hunting rabbit. So the rabbit hunters will have more offspring, which makes it even worse to hunt stags. Eventually, everyone will hunt rabbits.

The two outcomes (Stag,Stag) and (Rabbit,Rabbit) are the two Nash Equilibria in the Stag Hunt. Evolutionary game theory predicts that the proportion of stag and rabbit hunters in a population will approach one of these equilibria.

Not every Nash Equilibrium is a possible end point of evolution though. For example, if a population repeatedly plays the game of Chicken, and the players can't recognize in advance who will swerve and who will go straight, then the asymmetric equilibria (Swerve, Straight) and (Straight, Swerve) do not mark possible end points of evolutionary dynamics. But note that in a community in which almost everyone swerves, you're better off going straight; similarly, in a community in which almost everyone goes straight, the best choice is to swerve. Evolution will therefore lead to the third, mixed strategy equilibrium, which represents a state in which a certain fraction of the population swerves and the others go straight.

The assumption that individuals in a population are randomly paired with one another is obviously unrealistic. In reality, individuals are more likely to interact with members of their own family, which increases the chances that they will be paired with individuals of the same type; they might also actively seek out others who share the relevant traits. Either way, the resulting **correlated play** dramatically changes the situation.

For example, consider a population in which individuals repeatedly play a Prisoner Dilemma wherein they can either cooperate with each other (remain silent, in the original scenario) or defect (confess). Since defectors always do better than cooperators in any encounter, it may seem that cooperation can never evolve. However, cooperators do much better when paired with other cooperators than defectors when paired with defectors. If the extent of correlation is sufficiently high, cooperators can therefore take over (although perhaps not completely).

In many species, one can find altruistic individuals who sacrifice their own fitness for the sake of others. Evolutionary game theory explains how this kind

of altruism could have evolved.

Exercise 10.10 ⋆

Why can't we expect cooperative behaviour to take over completely in the scenario where cooperation spreads through correlated play?

Exercise 10.11 ⋆

What are the Nash equilibria in the following game (ignoring randomization)? Could all the equilibria come about through an evolutionary process?

	A	B
A	5,5	1,1
B	1,1	1,1

10.7 Further reading

The Stanford Encyclopedia entry on game theory provides a fairly comprehensive overview:

- Don Ross: "Game Theory" (2014)

The paradox of backward induction is discussed in

- Philip Pettit and Robert Sugden: "The Backward Induction Paradox" (1989)

For a little more on evolutionary game theory, see

- Brian Skyrms: "Game Theory, Rationality and Evolution of the Social Contract" (2000)

Essay Question 10.1

Explain the paradox of backward induction. Why is it a paradox? How do you think it could be resolved?

11 Bounded Rationality

11.1 Models and reality

We have studied an abstract model of rational agents. The model assumes that the agents have some idea of what the world might be like, which we represent by a credence function Cr over a suitable space of propositions. We also assume that the agents have some goals or values or desires, which we represent by a utility function U on the same space of propositions. An agent's credence function is assumed to satisfy the formal rules of the probability calculus; it is assumed to evolve over time by conditionalizing on sensory information, and it is assumed to satisfy further constraints such as the Probability Coordination Principle. The agent's utility function U is assumed to satisfy Jeffrey's axiom, so that it is jointly determined by the agent's credences and the utility assigned to basic concerns. These utilities may in turn be determined by aggregating subvalues. When the agent faces a choice, she is assumed to choose an act which maximizes the credence-weighted average of the utility of the act's possible outcomes.

There are different ways of filling in the details, so our model is really a family of models. Should expected utility be understood causally or evidentially? Should credences satisfy a restricted version of the Principle of Indifference? Should utilities conform to localism? Should they be stationary? Different answers yield different models.

Each model in this family can be understood either **normatively** or **descriptively**. Understood normatively, the model would purport to describe an ideal which real agents should perhaps aspire to. Understood descriptively, the model would purport to describe the attitudes and choices of ordinary humans.

It is a commonplace in current economics and psychology that our models are descriptively inadequate: that real people are not expected utility maximizers. In itself, this is not necessarily a problem – not even for the descriptive interpretation of our model. Remember that "all models are wrong". With the possible exception of the standard model of particle physics, the purpose of a model is generally to

identify interesting and robust patterns in the phenomena, not to get every detail right. Nonetheless, it is worth looking at how real agents fail to conform to our model, and what we could change to make it more realistic.

Many examples of people supposedly failing to maximize expected utility are not really counterexamples to the descriptive adequacy of our model, since the examples rely on implausible restrictions on the agent's utility function. As we saw in chapter 8, the expected-utility approach can easily accommodate agents who care about risk or fairness. We can similarly accommodate altruistic behaviour (section 1.2), the endowment effect (section 5.2), or apparent failures of time consistency (section 7.4).

But other phenomena are harder to accommodate. For example, suppose I offer you £100 for telling me the prime factors of 82,717. You have 10 seconds. Can you do it? Probably not. All you'd have to do to get the money is utter '181 and 457', which is surely an available act. Moreover, that '181 and 457' is the correct answer logically follows from simpler facts of which you are highly confident. By the rules of probability, you should therefore be confident that '181 and 457' is the correct answer. As an exptected utility maximizer (assuming you'd like to get the £100), you would utter these words. Yet you don't.

Exercise 11.1 ⋆⋆

Explain why, if some proposition C logically follows from two other propositions A and B, and $\mathrm{Cr}(A) > 0.9$ and $\mathrm{Cr}(B) > 0.9$, then $\mathrm{Cr}(C) > 0.81$.

In 1913, Ernst Zermelo proved that in the game of chess, there is either a strategy for the starting player, White, that guarantees victory no matter what Black does, or there is such a strategy for Black, or there is a strategy for either player to force a draw. Consequently, if two ideal Bayesian agents sat down to a game of chess, and their only interest was in winning, they would agree to a draw or one of them would resign before the first move. Real people don't play like that.

Another respect in which real people plausibly deviate from our model is that they often overlook certain options. You go to the shop, but forget to buy soap. You walk along the highway because it doesn't occur to you that you could take the nicer route through the park. The relevant options (buying soap, taking the nicer route) are available to you, and they are better by the lights of your beliefs and desires, so it is a mistake that you don't choose them.

Relatedly, real people are forgetful. I don't remember what I had for breakfast

last Monday. As an ideal Bayesian agent, I would still know what I had for breakfast on every day of my life.

Exercise 11.2 ⋆⋆

Show that if $\mathrm{Cr}_t(A) = 1$, and the agent conditionalizes on information E with $\mathrm{Cr}_t(E) > 0$, then $\mathrm{Cr}_{t+1}(A) = 1$. (Conditionalization was introduced in section 4.2.)

So we should admit that our model does not fit real agents in every respect. There is indirect evidence for this from research on artificial intelligence. The type of model we have studied is well known in these quarters, and forms the background for much recent research. Yet the model turns out to be computationally intractable. Real agents with limited cognitive resources, it seems, couldn't possibly conform to our model.

11.2 Avoiding computational costs

Before we look at ways of making our model more realistic, I want to address another common misunderstanding.

Suppose you walk back to shop to buy soap. At any point on your way, you could decide to turn around, or start running, or check if your shoe laces are tied, or mentally compute $181 \cdot 457$, or start humming the national anthem, or utter 'Age quod agis', and so on. There are millions of things you could do. Many of these would lead to significantly different outcomes, especially if you consider long-term consequences. (Hitler almost certainly would not have existed if hours or even months before his conception, his mother had decided to run rather than walk to buy soap.) Some people take the MEU Principle to imply that at each point of your walk to the shop, you should explicitly consider all your options, envisage all their possible outcomes, assess their utility and probability, and on that basis compute their expected utility. This is clearly unrealistic and infeasible.

But the MEU Principle requires no such thing. The MEU Principle says that rational agents choose acts that maximize expected utility; it specifies *which acts* an agent should choose, given her beliefs and desires. It says nothing about the internal processes that lead to these choices. It does not say that the agent must explicitly consider all her options and compute expected utilities.

Exercise 11.3 ⋆⋆

The opposite is closer to the truth. Suppose an agent has a choice between turning left (L), turning right (R), and sitting down to compute the expected utility of L and R and then choosing whichever comes out best. Let C be that third option. If computing expected utilities involves some costs in terms of effort or time, then either L or R generally has greater expected utility than C. Explain why.

The MEU Principle does not require calculating expected utilities. But that raises a puzzle. An agent who conforms to our model always chooses acts with greatest expected utility. How is she supposed to do that without calculating? It doesn't seem rational to choose one's acts randomly and maximize expected utility out of sheer luck.

Part of the answer is that in many circumstances, simple alternatives to computing expected utilities reliably lead to optimal choices. As the psychologist Gerd Gigerenzer once pointed out, if you want to catch a flying ball, an efficient alternative to computing the ball's trajectory – which is generally intractable – is to move around in such a way that the angle between you and the ball remains within a certain range. This ensures that you'll eventually stand where the ball will arrive. If you desire to catch the ball, following Gigerenzer's heuristic will maximize expected utility. You don't need to consciously compute anything, and you don't need to conceptualize what you're doing as maximizing expected utility.

Exercise 11.4 ⋆

Suppose you're a musician in the middle of a performance. Trying to compute the expected utility of all the notes you could play next would probably derail your play. Even if it wouldn't, it would completely change your experience of playing, probably for the worse. Give another example where conceptualizing one's acts as maximizing expected utility would undermine the value of performing the acts.

One reason why many decision problems don't require sophisticated computations is that a certain act clearly dominates all the others. Whether that is the case depends on the agent's utility function. It follows that you can reduce the computational costs of decision-making by tweaking your utilities. For example, suppose you assign significant (sub)value to obeying orders. Doing whatever you're ordered to do is then a reliable way of maximizing expected utility, and

it requires very little cognitive effort. Similarly if you value imitating whatever your peers are doing.

Our capacity for planning and commitment can also be seen in this light. Before you went to the shop, you probably decided to go to the shop. The direct result of that decision was an intention to go to the shop. Once an intention or plan is formed, we are motivated to executing the plan. Revising a plan or overturning a commitment has negative (sub)value. Consequently, once you've formed an intention, simply following it reliably maximizes expected utility. You don't need to think any more about what to do unless you receive surprising new information or your basic values suddenly change. (This is true even if you made a mistake when you originally formed the intention.)

Habits can play a similar role. Most of us spend little effort deciding whether or not to brush our teeth in the morning; we do it out of habit. Habitual behaviour is computationally cheap, and it can reliably maximize expected utility – especially if we assign (sub)value to habitual behaviour. And we do, at least on a motivational conception of desire: habits motivate.

The upshot is that various cognitive strategies that are often described as alternatives to computing expected utilities – habits, instincts, heuristics, etc. – may well be efficient techniques for maximizing expected utility. Far from ruling out such strategies, our model actually predicts that we should use them.

An example in which something like this might play a role is Ellsberg's Paradox, another classical "counterexample" to the MEU Principle.

Example 11.1 (Ellsberg's Paradox)

An urn contains 300 balls. 100 of the balls are red, the others are green or blue, in unknown proportion. A ball is drawn at random from the urn. Which of the following two gambles do you prefer?

	Red	Green	Blue
A	£1000	£0	£0
B	£0	£1000	£0

Next, which of C and D do you prefer?

	Red	Green	Blue
C	£1000	£0	£1000
D	£0	£1000	£1000

Many people prefer A to B and D to C. Like in Allais's Paradox, there is no way of assigning utilities to the monetary outcomes that supports these preferences.

Exercise 11.5 ⋆

Assume the outcomes in Ellsberg's paradox are described correctly and you prefer more money to less. By the Probability Coordination Principle, $\text{Cr}(\textit{Red}) = {}^1/_3$. What would your credences in *Green* and *Blue* have to look like so that $EU(A) > EU(B)$? What would they have to look like so that $EU(D) > EU(C)$?

In Ellsberg's Paradox, risk aversion doesn't seem to be at issue. What makes the difference is that you know the objective probability of winning for options A and D: it is ${}^1/_3$ for A and ${}^2/_3$ for D. But you don't know the objective probability of winning with B and C, since you have too little information about the non-red balls.

Why does that matter? One explanation is that people simply prefer lotteries (in which the outcomes have known objective probabilities) to uncertain prospects (in which only subjective probability can be given to the outcomes). With such a utility function, the outcome wrongly labelled '£1000' in A is actually better than the corresponding outcome in C, because only the former involves having chosen a lottery.

Exercise 11.6 ⋆

The explanation of the Ellsberg preferences that I just outlined makes the preferences conform to the MEU Principle by redescribing the outcomes. Is the redescription global or local in the sense of chapter 8?

But why would agents prefer lotteries? Perhaps because such a preference can reduce computational costs. If you know the objective probabilities of the states, it is easy to figure out the credence you should give to the states: it should match the objective probabilities. If you don't know the objective probabilities, a lot more work may be required to figure out the extent to which your total evidence supports the various states. In Ellsberg's Paradox, Cr(Red) is a easier to figure out than Cr(Green) and Cr(Blue). If you have a preference for lotteries, you don't need to figure out Cr(Green) and Cr(Blue): from eyeballing the options, you can already see that the expected monetary payoff of A and B is approximately the same (ditto for C and D); a preference for lotteries then clearly favours A (and D).

11.3 Reducing computational costs

I will now review a few ideas from theoretical computer science for rendering our models computationally tractable.

Imagine we are designing an artificial agent, with a probabilistic representation of her environment and a number of goals or desires. Let's assume we want the agent to have credences and utilities in 50 logically independent propositions $A_1, \ldots, A_{50}$ (an absurdly small number). How large of a database do we need for that?

You might think that we need 50 records for the probabilities and 50 for the utilities. But we generally can't compute $Cr(A \wedge B)$ or $\mathrm{Cr}(A \vee B)$ from $\mathrm{Cr}(A)$ and $\mathrm{Cr}(B)$. Nor can we compute $U(A \wedge B)$ or $U(A \vee B)$ from $U(A)$ and $U(B)$. To be able to determine the agent's entire credence and utility functions (without further assumptions), we need to store at least the probability and utility they assign to every "possible world" – that is, to every maximally consistent conjunction of $A_1, \ldots, A_{50}$ and their negations.

Exercise 11.7 ⋆⋆⋆

Explain why the probability of every proposition that can be defined in terms of $A_1, \ldots, A_{50}$ can be computed from the probability assigned to the possible worlds. Then explain why the utility of all such propositions can be computed from the probability and utility assigned to the worlds.

There are $2^{50} = 1,125,899,906,842,624$ maximally consistent conjunctions of $A_1, \ldots, A_{50}$ and their negations. Since we need to store credences and utilities, we therefore need a database with $2,251,799,813,685,248$ records. (I'm exaggerating a little. Once we've fixed the probability of the first 1,125,899,906,842,623 worlds, the probability of the last world is 1 minus the sum of the others, so we really only need $2,251,799,813,685,247$ records.)

We'll need to buy a lot of hard drives for our agent if we want to store 2 quadrillion floating point numbers. Worse, updating all these records in response to sensory information, or computing expected utilities on their basis, will take a very long time, and use a large amount of energy.

In chapter 7, we've encountered two tricks for simplifying the representation of an agent's utility function. First, if the agent cares about some features of the world and not about others, it is enough to store the agent's utility for her "concerns": the maximally consistent conjunctions of the features she cares about

(section 5.4). For example, if our agent only cares about the possible combinations of 20 among the 50 propositions $A_1, \ldots, A_n$, we only need to store 2^{20} values. Second, and more dramatically, if the agent's preferences are separable, we can further cut down the number of utility records from 2^{20} to $2 \cdot 20 = 40$, because the value of any combination of the 20 propositions and their negations can be determined by adding up the relevant subvalues (section 7.2).

Similar tricks are available for the agent's credence function. Mirroring the first trick, we could explicitly store only the agent's credence in certain sets of possible worlds, and assume that her credence is distributed uniformly within these sets. The trick can be extended to non-uniform distributions. For example, suppose our agent has imperfect information about how far she is from the next charging station. Instead of explicitly storing a probability for every possible distance (1 m, 2 m, 3 m, ...), we might assume that the agent's credence over these possibilities follows a Gaussian (or binomial) distribution, which can be specified by two numbers (mean and variance). Researchers in artificial intelligence make heavy use of this trick.

An analogue of separability, for credences, is probabilistic independence. If A and B are probabilistically independent, then $\mathrm{Cr}(A \wedge B) = \mathrm{Cr}(A) \cdot \mathrm{Cr}(B)$. If all the 50 propositions $A_1, \ldots, A_{50}$ are mutually independent, then we can fix the probability of all possible worlds and therefore of all logical combinations of the 50 propositions by specifying their individual probability.

Independence is often plausible. Whether the next charging station is 100 meters away plausibly doesn't depend on whether the outside temperature is above 20°C. But for many other propositions, independence is implausible. On the supposition that it is warm outside, it may well be more likely that the window is open, or that there are people on the street, than on the supposition that it isn't warm. If the agent is unsure whether it is warm, it follows that $\mathrm{Cr}(\mathit{Open}/\mathit{Warm}) > \mathrm{Cr}(\mathit{Open})$, and $\mathrm{Cr}(\mathit{People}/\mathit{Warm}) > \mathrm{Cr}(\mathit{People})$. So we can't assume probabilistic independence across all the 50 propositions $A_1, \ldots, A_{50}$.

Even where independence fails, however, we often have **conditional independence**. For example, if warm temperatures make it more likely that the window is open and that there are people on the street, then an open window is also evidence that there are people on the street: $\mathrm{Cr}(\mathit{People}/\mathit{Open}) > \mathrm{Cr}(\mathit{People})$. So *People* and *Open* are not independent. However, *on the supposition that it is warm outside*, the window being open may no longer increase the probability of people

on the street:

$$\mathrm{Cr}(\mathit{People}/\mathit{Open} \wedge \mathit{Warm}) = \mathrm{Cr}(\mathit{People}/\mathit{Warm}).$$

In that case, we say that *People* and *Open* are independent *conditional on Warm*.

Now consider the possible combinations of *Warm*, *People*, *Open* and their negations. By the probability calculus (compare exercise 2.10),

$$\mathrm{Cr}(\mathit{Warm}\wedge\mathit{People}\wedge\mathit{Open}) = \mathrm{Cr}(\mathit{Warm})\cdot\mathrm{Cr}(\mathit{Open}/\mathit{Warm})\cdot\mathrm{Cr}(\mathit{People}/\mathit{Open}\wedge\mathit{Warm}).$$

By the above assumption of conditional independence, this simplifies to

$$\mathrm{Cr}(\mathit{Warm} \wedge \mathit{People} \wedge \mathit{Open}) = \mathrm{Cr}(\mathit{Warm}) \cdot \mathrm{Cr}(\mathit{Open}/\mathit{Warm}) \cdot \mathrm{Cr}(\mathit{People}/\mathit{Warm}).$$

In general, with the assumption of conditional independence, we can fix the probability of all combinations of *Warm*, *People*, *Open*, and their negations by specifying the probability of *Warm*, the probability of *People* conditional on *Warm* and on $\neg$*Warm*, and the probability of *Open* conditional on *Warm* and on $\neg$*Warm*. This reduces the number of required records from $2^3 - 1 = 7$ to 5, which may not look all that impressive, but the method really pays off if more than three propositions are involved.

The present technique of exploiting conditional independence to simplify probabilistic models is known under the heading of **Bayesian networks** (or **Bayes nets**, for short). Bayes nets have proved useful in wide range of applications.

A special case of Bayes nets is commonly used in artificial intelligence to model decision-making agents. A decision-making agent needs not only information about the present state of the world, but also about the future. We can model a whole history of states as a sequence $\langle S_1, S_2, S_3, \dots \rangle$, where S_1 is a particular hypothesis about the present state, S_2 about the next state, and so on. If there are 100 possible states at any given time, there will be $100^{10} = 100,000,000,000,000,000,000$ possible histories with length 10. Instead of storing individual probabilities for all these histories, it helps to assume that a later state (probabilistically) depends only on its immediate predecessor, so that $\mathrm{Cr}(S_3/S_1 \wedge S_2) = \mathrm{Cr}(S_3/S_2)$. This is known as the **Markov assumption**. It reduces the number of records we'd need to store from 100^{10} to 990,100.

To further simplify the task of decision-making, computer scientists typically assume that basic values are stationary and separable across times, so that the value of a history of states is a discounted sum of the subvalue for individual states.

To specify the whole utility function, we then only need to store the discounting factor δ and 100 values for the individual states. The task of conditionalization can also be simplified, by assuming that sensory evidence only contains direct information about the present state of the world, rather than entire histories.

These simplifications define what computer scientists call a '**POMDP**': a **Partially Observable Markov Decision Process**. There is a simple recursive algorithm for computing expected utilities in POMDPs.

In practice, even these simplifications generally don't suffice to make conditionalization and expected utility maximization tractable. Further simplifications are needed. For example, it often helps to "myopically" ignore states in the distant future and let the agent maximize the expected utility for the next few states only. In addition, various techniques have been developed that allow an efficient *approximate* computation of expected utilities and posterior probabilities. Such techniques are often supplemented by a meta-decision process which lets the system choose a level of precision: when a lot is at stake, it is worth spending more effort on getting the computations right.

While originating in theoretical computer science, these models and techniques have in recent years had a great influence on our models of human cognition. There is evidence that when our brain processes sensory information or decides on a motor action, it employs the same techniques computer scientists have found useful in approximating the Bayesian ideal. Several quirks of human perception and decision-making can apparently be seen as consequences of the shortcuts our brain uses to approximate conditionalization and computing expected utilities.

11.4 "Non-expected utility theories"

Meanwhile, researchers at the intersection of psychology and economics have also tried to develop more realistic models of decision-making. The most influential of these alternatives is **prospect theory**, developed by Daniel Kahneman and Amon Tversky.

Prospect theory has to be understood on the background of a highly restricted version of decision theory that dominates economics. The highly restricted theory only deals with choices between lotteries, with known objective probabilities. The outcomes of these lotteries are identified with monetary wealth or commodity bundles. (Even local feelings of frustration or regret are excluded.) Finally, it is assumed that agents always prefer more money or goods, with declining marginal

utility. When you find social scientists discuss "Expected Utility Theory", this highly restricted theory is what they usually have in mind. Prospect theory now proposes four main changes.

1. *Reference dependence*. According to prospect theory, agents classify possible outcomes into gains and losses, by comparing the outcomes with a contextually determined reference point. Outcomes better than the reference point are modelled as having positive utility, outcomes worse than the reference point have negative utility.

2. *Diminishing sensitivity*. Prospect theory holds that both gains and losses have diminishing marginal utility: the same objective difference in wealth makes a larger difference in utility near the reference point than further away, on either side. For example, the utility difference between a loss of £100 and a loss of £200 is greater than that between a loss of £1000 and a loss of £1100. This predicts that people are risk averse about gains but risk seeking about losses: they prefer a sure gain of £500 to a 50 percent chance of £1000, but they prefer a 50 percent chance of losing £1000 to losing £500 for sure.

3. *Loss aversion*: According to prospect theory, people are more sensitive to losses than to gains of the same magnitude. For example, the utility difference between a loss of £100 and a loss of £200 is greater than that between a gain of £200 and a gain of £100. This explains why many people turn down a lottery in which they can either win £110 or lose £100, with equal probability.

4. *Probability weighting*. According to prospect theory, the outcomes are weighted not by their objective probability, but by transformed probabilities known as 'decision weights' that are meant to reflect how seriously people take the relevant states in their choices. Decision weights generally overweight low-probability outcomes. Thus probability 0 events have weight 0, probability 1 events have weight 1, but in between the weight curve is steep at the edges and flatter in the middle: probability 0.01 events might have weight 0.05, probability 0.02 events weight 0.08, ..., probability 0.99 events weight 0.92. Among other things, this is meant to explain why people play the lottery, and why they tend to pay a high price for certainty: they prefer a settlement of £90000 over a trial in which they have a 99% chance of getting £100000 but a 1% chance of getting nothing.

Prospect theory is clearly an alternative to the simplistic economical model mentioned above. It is not so obvious whether it is alternative to the more liberal model we have been studying for most of this course. Diminishing sensitivity and loss aversion certainly don't contradict our model. Reference dependence and

probability weighting are a little more subtle.

Our model assumes that if an agent knows the objective probability of a state, then in decision-making she will weight that state in proportion to the known probability. Prospect theory says that real people don't do that. If we measure an agent's credences in terms of preferences or choices, then the decision weights of prospect theory are the agent's credences: they play precisely the role of credences in guiding behaviour. So prospect theory assumes that people systematically violate the Probability Coordination Principle, since their credence in low-probability event is greater than the known objective probability.

Some have argued that the observations that motivate probability weighting are better explained by redescribing the outcomes and allowing people to care about things like risk or fairness. However, there is evidence that sometimes people really do fail to coordinate their beliefs with known objective probabilities, especially if the probabilities are communicated verbally. – Studies show that people's decision weights are closer to the objective probabilities if they have experienced these probabilities as relative frequencies in repeated trials. By contrast, when people reason explicitly about probabilities, systematic mistakes like the base rate fallacy are very common.

Reference dependence may also raise a genuine challenge. To be sure, most forms of reference dependence are harmless. Our model can easily accommodate people who care especially about how much they will have in comparison to what they had before, or in comparison to what their peers have. But sometimes, the reference point is affected by intuitively irrelevant features of the context, and that is harder to square with our model.

Exercise 11.8 ⋆

When people compete in sports, average performance sometimes seems to function as a reference point, insofar as the effort people put in to avoid performing below average is higher than the effort they put in to exceed the average. Can you explain this observation by "redescribing the outcomes" in the model we have studied, without appealing to reference points?

The problematic type of reference dependence is closely related to so-called **framing effects**. In experiments, people's choices can systematically depend on how one and the same decision problem is described. For example, when presented with a hypothetical situation in which 1000 people are in danger of death, and a certain act would save exactly 600 of them, subjects are more

favourable towards the act if it is described in terms of '600 survivors' than if it is described in terms of '400 deaths'. In prospect theory, the difference might be explained by a change in reference point: if the outcome is described in terms of survivors, it is classified as a gain; if it is described in terms of deaths, it is classified as a loss.

In principle, our liberal model could also explain the relevance of the description. Perhaps people assign basic value to choosing options *that have been described in terms of survivors* rather than in terms of deaths. However, on reflection, most people would certainly deny that the verbal description of an outcome is of great concern to them. As in the case of decision weights, a more adequate model would arguably have to take into account our incomplete grasp of a verbally described scenario. When hearing about survivors, we focus on a certain attribute of the outcome, on all the people who are saved. That attribute is desirable. When hearing about deaths, a different, and much less desirable, attribute of the same outcome becomes salient.

Ideal agents always weigh up all attributes of every possible outcome. Real agents arguably don't do that, as it requires considerable cognitive effort. As a result, the attributes we consider depend on contextual clues such as details of a verbal description. Some recent models of decision making in philosophy and psychology take this kind of attribute selection into account.

11.5 Imprecise credence and utility

I'm going to toss three dice. Would you rather get £1000 if the sum of the numbers on the dice is at least 10 or if all three numbers are different? You'd probably need some time to give a final answer. You know that the six possible results for every dice have equal probability, and that the results are independent. But it takes some effort to infer from that which of the two events I described is more likely.

A real agent's cognitive system can't explicitly store her credence and utility for every proposition. It can only store a limited number of **constraints** on credences and utilities. A constraint rules out some credences and utilities, but not others. For example, that outcomes of die tosses are probabilistically independent is a constraint; among other things, it entails that the probability of three sixes is the product of the probabilities for the individual dice: $\mathrm{Cr}(\mathrm{Six}_1 \wedge \mathrm{Six}_2 \wedge \mathrm{Six}_3) = \mathrm{Cr}(\mathrm{Six}_1)\mathrm{Cr}(\mathrm{Six}_2)\mathrm{Cr}(\mathrm{Six}_3)$, but it does not fix what these probabilities are.

So far, we have assumed that taken together, the constraints stored by an agent's cognitive system are rich enough to determine a unique credence and utility function. But maybe they don't. Maybe there are questions on which you don't have a settled opinion, even in principle, after ideal reflection. Or suppose you don't have time for lengthy reflection, or you're too tired. In such cases, it would arguably be wrong to model your attitudes in terms of a single, precise credence function, and a single, precise utility functions.

Across several disciplines, researchers have developed models which relax the assumption of unique and precise credences and utilities. The standard approach is to use **sets of credence and utility functions** instead of single functions. The functions in the set are all those that meet the constraints. Intuitively, each member of the set is a *refinement* or *precisification* of the agent's indeterminate state of mind.

The use of sets of credence functions is often motivated by consideration like the following. How likely do you think it is that it will snow in Toronto on 7 January 2041? If someone suggested the probability is 80%, you might say that's too high; 5% seems too low. But 20% might seem just as plausible to you as 21%. It would therefore be wrong to model your state of mind by single and precise probability. Rather, we should say that your credence is a whole range of numbers, like so:

$$\text{Cr}(\mathit{Snow}) = [0.1, 0.5].$$

Here, '$[0.1, 0.5]$' denotes the range of all numbers from 0.1 to 0.5; it is the range of all numbers your refined credence functions assign to *Snow*.

You should be skeptical about this line of argument. The term 'probability' in English almost always means objective probability. When asked about the probability of *Snow*, it is therefore natural to interpret the question as concerning a certain objective quantity – something you could perhaps find out by developing a sophisticated weather model. But credence, on the Bayesian conception, is not belief about objective probability. It is simply strength of belief. You could not find out your credence in *Snow* by developing a sophisticated weather model.

Nonetheless, there are reasons to extend the Bayesian conception of credence to allow for sets of credence functions. For example, suppose we measure (or define) an agent's credences and utilities in terms of her preferences. Various representation theorems show that if the agent's preferences satisfy certain axioms, then the preferences are represented by a *unique* credence and utility function (except for the conventional choice of zero and unit for utilities). Giving up

uniqueness then means that the agent violates one or more of the axioms. And it is not implausible that real agents do violate some of these axioms.

In particular, consider the completeness axiom. Completeness states that for any propositions A and B, the agent either prefers A to B or B to A or is indifferent between the two. This is trivial if we define preference in terms of choice. Indeed, presented with a forced choice between A and B, you will inevitably choose either A or B; even indifference can be ruled out. But we've already seen that if we want to measure credence and utility in terms of preference, then the relevant preference relation can't be directly defined in terms of choices. And once we take a step back from choice behaviour, it seems perfectly possible that you might neither prefer A to B, nor B to A, and yet you're not indifferent between the two. Instead, you haven't fully made up your mind. The two propositions seem roughly "on a par", but you wouldn't say they are exactly equal in value.

For example, would you rather lose your capacity to hear or your capacity to walk? You may well have no clear preference, even after considerable reflection. Does that mean you're exactly indifferent? Not necessarily. If you were, you should definitely prefer losing the capacity to hear *and getting £1* to losing the capacity to walk. In reality, the added £1 probably doesn't make a difference.

Exercise 11.9 ⋆⋆

Suppose we define '$\sim$' in terms of '$\succ$', as follows: $A \sim B \iff (A \not\succ B) \wedge (B \not\succ A)$. Completeness is then logically guaranteed. But other assumptions about preference then fail if you haven't made up your mind between certain propositions. For example, it is possible to have $A \succsim B$, $B \succsim C$, and $C \succsim A$, contradicting transitivity. Explain why (assuming that $A \succsim B \iff (A \succ B) \vee (A \sim B)$).

So there are reasons to relax completeness, at least if we're interested in modelling real agents. (Some would say even ideal agents don't need to have complete preferences.) But we may still impose an axiom of **completability**. That is, we can require that if an agent's preferences violate, say, the Savage axioms because they fail to rank certain options, then there is a refinement of her preferences, filling in the missing rankings, that does satisfy the axioms. Savage's representation theorem then implies that the agent's preferences are represented by a set of credence and utility functions.

In sum, even if we don't conflate credence with belief about objective probability, we might want to model agents as having a set of credence (and utility) functions. We then have to revise other parts of our model, to explain how those agents

should update their beliefs over time, and how they should make choices. As it turns out, the required revisions are not at all straightforward. I will mention just one problem, concerning rational choice.

Suppose you have a set of credence and utility functions, because you haven't made up your mind about certain things, and you face a choice. According to some of your credence and utility functions, act A has greatest expected utility; according to others, you should choose B. What should you do? A popular "permissivist" answer is that you are permitted to choose either option.

Exercise 11.10 ⋆⋆

Explain how the preference of A over B and D over C in Ellsberg's paradox might be justified by the permissivist approach, without redescribing the outcomes. (What is the expected utility of the four options?)

But now consider the following scenario. You are offered two bets A and B, one after the other, on a proposition H about which you haven't made up your mind. Let's say $\text{Cr}(H) = [0.2, 0.8]$. Bet A would give you £1.40 if H and £-1 if $\neg H$. Bet B would give you £-1 if H and £1.40 if $\neg H$. Assuming for simplicity that your utility is precise and proportional to the monetary payoff, both bets have an imprecise expected utility of between -0.52 and 0.92. (For example, the expected utility of the first bet ranges from $0.2 \cdot -1 + 0.8 \cdot 1.40 = 0.92$ to $0.8 \cdot -1 + 0.2 \cdot 1.40 = -0.52$.) Your undecided state of mind therefore leaves open whether accepting either bet is a good idea. On the permissivist approach, it is permissible for you to refuse both bets. But arguably, that would be irrational, since the two bets together have a guaranteed payoff of £0.40. By refusing both bets, you would miss out on a sure gain.

11.6 Further reading

An accessible overview of some advances in theoretical computer science and their influence on cognitive science is

- Samuel Gershman et al: "Computational rationality: A converging paradigm for intelligence in brains, minds, and machines" (2015)

For a brief overview of prospect theory and related models, motivated by the idea of bounded rationality, see

- Daniel Kahneman: "A Perspective on Judgment and Choice" (2003)

The Stanford Encyclopedia article on imprecise probabilities gives a thorough (and highly opinionated) overview of the models discussed in section 11.5:

- Saemus Bradley: "Imprecise Probabilities" (2014)