Belief, Desire, and Rational Choice

Wolfgang Schwarz

December 20, 2022

github.com/wo/bdrc

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 Miners 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 have enough sandbags to block one shaft, but not both. If you block the right shaft, all miners will survive. If you block the wrong shaft, all of them will die. If you do nothing, both shafts will fill halfway with water and one miner (the shortest of the ten) will die.

What should you do?

There's a sense in which the answer depends on where the miners are. If they are in shaft A then 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 not what you should do *in light of all the facts*, but what you should do *in light of your information*. 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 high 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 do block neither shaft and 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 situation, 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 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.

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. There's a sense in which the doctor's recommendation was wrong, and a sense in which it was right. 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 Miners Problem, there are three acts: block shaft *A*, block shaft *B*, and block neither shaft. ('Possible act' would be a better name: if, say, you decide to block shaft *B*, 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 Miners 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 acts leads to one of the outcomes, but the decision-maker often doesn't know how the outcomes are associated with the acts. In the Miners Problem, 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.

The dependency between acts and outcomes is captured by the **states**. Informally, a state specifies the external circumstances that determine which choice would lead to which outcome. The Miners Problem has two relevant states: that the miners are in shaft *A*, and that they 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 Miners Problem in a table, called a **decision matrix**:

	Miners in *A*	Miners in *B*
Block *A*	all 10 live	all 10 die
Block *B*	all 10 die	all 10 live
Block neither	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's Dilemma)

You and your partner have been arrested for some crime and are separately interrogated. If you both confess, you will each 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 serve one year each.

The Prisoner's Dilemma combines two decision problems: one for you and one for your partner. We could also think about a third problem that you face as a group. But 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 a better outcome no matter what your partner does.

I've assumed that you only care about your own prison term. What if you also care about your partner's fate? Then your decision problem is not adequately summarized by the above matrix, because the cells in the matrix don't say what happens to your partner. The "outcomes" in a decision problem must specify everything that matters to the agent. If you care about your partner, the matrix might look as follows.

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

Now confessing is no longer the obviously best choice. If, for example, your aim 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.

Exercise 1.3 †††

In an adequate decision matrix, the states must be independent of the acts: which state obtains should not be affected by which act is chosen. The following decision matrix was drawn up by a student who wonders whether to study for an exam. It suggests that not studying is guaranteed to lead to a better outcome. However, the matrix violates the independence requirement. Can you draw an adequate matrix for the student's decision problem?

	Will Pass	Won't Pass
Study	Pass & No Fun	Fail & No Fun
Don't Study	Pass & Fun	Fail & Fun

1.3 Belief, desire, and degrees

To solve a decision problem we generally need to know what the agent wants and what she believes. Typically, we also need to know *how strong* these attitudes are.

Return to the Mushroom Problem. Suppose you like eating a delicious mushroom, and you dislike being hungry and being dead. We might 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. 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. We will once again represent the valence and strength of beliefs by numbers, but this time we only use numbers between 0 and 1. If an agent is certain that a given state obtains, then her degree of belief in that state 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 that it involves a measure of (rational) degree 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 your credence in the paddy straw possibility and your credence in the death cap possibility add up to less than 1. But then our decision matrix should include further columns for the other possibilities.

Are there also formal constraints on rational degrees of desire? This is less obvious. The fact that your utility for eating a paddy straw is +1, for example, does not seem to entail anything about your utility for eating a death cap. In later chapters, we will see that utilities nonetheless have an interesting 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 entirely self-interested. Accordingly, 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 1.4 †

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 the ingredients 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's Dilemma, given that all you care about is minimizing your own prison term. The fully labelled matrix might look as follows.

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

Since confessing leads to a better outcome no matter what your partner does, it is obviously the right choice. We don't even need to look at what you think your partner will do.

An act that leads to a better outcome than another in every state is said to **dominate** the other act. An act that dominates all other acts is called **dominant**. For agents who only care about themselves, confessing is the dominant option in the Prisoner's Dilemma.

The Prisoner's Dilemma is famous because it refutes the idea that good things will always come about if people only look after their own interests. If the two parties in the Prisoner's Dilemma want to minimize their own prison term, 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. 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. As a general decision rule, however, it 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 that 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 evaluates each act by the *weighted average* of the utility of all outcomes the act might bring about, weighted by the probability 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, …, x_n, then their average is

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

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

Let's compute the weighted average of the utility that might result from eating the mushroom in the Mushroom Problem. We multiply the utility of each outcome this act might bring about (+1 and -100) by your credence in the corresponding state (0.8 and 0.2), and then add up these products. The result is called the **expected utility** of

eating the mushroom.

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

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 '$\mathrm{U}(O_1)$' denote the utility of O_1, '$\mathrm{U}(O_2)$' the utility of O_2, etc. Then the expected utility of A is defined as

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

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

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

The term 'expected utility' is a little misleading. If you eat the mushroom in the Mushroom Problem, you are guaranteed to get either an outcome with utility +1 or an outcome with utility -100. You would not expect to get -19.2 units of utility. In the confusing lingo of probability theory, '**expectation**' simply means 'probability-weighted average'. The "expected outcome" of a die toss, for example, is

$$1/6 \cdot 1 + 1/6 \cdot 2 + 1/6 \cdot 3 + 1/6 \cdot 4 + 1/6 \cdot 5 + 1/6 \cdot 6 = 3.5,$$

assuming all six outcomes have probability $1/6$. Here, too, it would be odd to literally expect the outcome 3.5.

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

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

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 1.5 †

Put (sensible) utilities and credences into the decision matrix for the Miners Problem, and compute the expected utility of the three acts.

Exercise 1.6 ††

Explain why the following decision rule is not generally reasonable: *Choose an act that leads to the best outcome in the most likely state (or in one of the most likely states, if there is a tie).*

Exercise 1.7 †††

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

Exercise 1.8 ††

Is this correct? *If an act is certain not to bring about the best outcome, then it should not be chosen.*

In the Mushroom Problem, the MEU Principle says that you shouldn't eat the mushroom. Although the most likely outcome of eating the mushroom has a positive utility, the expected utility of eating the mushroom is -19.2. A really good or really bad outcome can have a big impact on an act's expected utility even if the outcome is very improbable.

This effect is easy to miss. It is tempting to think, for example, that avoiding a plane trip in order to reduce one's carbon emissions is a pointless gesture: the plane isn't going to stay on the ground just because you don't take the trip. True. But

there is a chance that fewer flights will be scheduled in the future as a result of your choice. 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 tiny. On the other hand, the reduction in carbon emissions would be huge. *On average*, it has been estimated, a single person not taking a single flight reduces overall emissions by a little less than the flight's emissions divided by the number of seats on the plane. This is the "expected" effect of your choice, unless your case is unusual in other respects.

Even Nobel-price winning decision theorists are not immune to this kind of error. In 1980, John Harsanyi argued that utilitarian citizens who care only about the common good still have no good reason to participate in elections, given that any individual vote is almost certain not to make a difference. In one of his simplified examples, he assumes that a "very desirable policy measure M" gets implemented only if 1000 eligible voters all come to the polls and vote for it. Voting entails a minor cost in terms of convenience, but it would be better for everyone if the measure is passed than if (say) nobody votes and the measure isn't passed. Harsanyi claims that if the voters are rational then "each voter will vote only if he is reasonably sure that all other 999 voters will vote". Is this true?

Let's assume that each vote would decrease the overall welfare in the population by 1 unit (due the inconvenience for the voter). Since it would be better if everyone voted and the measure M were passed than if nobody voted and the measure fails, M must increase overall welfare by more than 1000. Now consider a utilitarian voter who only cares about overall welfare. If you do the math, you can see that voting maximizes expected utility for such a voter even if her credence that all the others will vote is as low as 0.001. She doesn't need to be "reasonably sure", as Harsanyi claims, that all the others will vote.

Exercise 1.9 †††

Do the math. 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.

Exercise 1.10 (Pascal's Wager) ††

One of the first recorded uses of the MEU Principle 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 1.11 ††

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

1.5 The problem of intentionality

A major obstacle to the systematic study of belief and desire is the apparent familiarity of the objects. We think and talk about beliefs and desires (our own, and other people's) from an early age, and continue to do so every day. We may sometimes ask how a peculiar belief or unusual desire came about, but the nature and existence of the states seems unproblematic. It takes effort to appreciate what philosophers call **the problem of intentionality**: the problem of explaining what beliefs and desires ultimately are.

To see the problem, assume (as many philosophers do) that people are nothing but large swarms of particles. What about such a swarm of particles could settle that it believes in, say, extraterrestrial life? Alternatively, ask yourself what we would have to do in order to create an artificial agent with a belief in extraterrestrial life. (Notice that it is neither necessary nor sufficient that the agent produces the sounds 'there is life on other planets'.)

If we allow for degrees of belief and desire, the problem of intentionality takes on a slightly different form. We need to explain what it ultimately means that an agent

has a belief or desire *with a particular strength*. What, exactly, do I mean when I say that my credence in extraterrestrial life is greater than 0.5, or that I give greater utility to sleeping in bed than to sleeping on the floor?

These may sound like obscure philosophical questions, but they are important for a proper assessment of the model we are going to study. There is a lot of cross-talk in the literature because different authors mean somewhat different things by 'credence' and 'utility'.

Conversely, it has been argued that the kind of model we will study holds the key to answering the problem of intentionality. Very roughly, the idea is that what it means to have such-and-such beliefs and desires is to act in a way that would make sense in light of these beliefs and desires.

I speak of beliefs and desires, but it might be better to stick with 'credence' and 'utility'. We should not assume that our ordinary psychological vocabulary precisely carves out the object of our investigation. The word 'desire', for example, can suggest 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. An agent's utilities, by contrast, are assumed to 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'.

Similar reservations apply to 'belief'. For example, some hold that one can have genuine beliefs only if one has a language (or "conceptually structured mental representations", whatever that is). We don't make any such assumption. Many animals have an inner representation of their environment that can be usefully modelled by a credence function, even though they don't have a language.

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 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. We won't pretend that we have a perfect grip on these quantities from the outset. Informally, an agent's credences capture how she takes the world to be, while her utilities capture how she

would like the world to be. We'll start with this vague and intuitive conception, and successively refine it 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. "All models are wrong", as the statistician George Box once put it. The aim of a model (outside fundamental physics) is not to provide a complete and fully accurate description of a certain aspect of reality – be it the diffusion of gases, the evolution of species, or the relationship between interest rates and inflation. The aim is to isolate simple and robust patterns in the relevant facts. 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.

Essay Question 1.1

Ordinary people arguably don't have fully precise and determinate degrees of belief. Suppose we model an agent's attitudes with an "imprecise" probability measure that assigns to each state a *range* of probabilities – 'between 0.2 and 0.4', for example. Can you find (and defend) a decision rule for agents of this kind?

Sources and Further Reading

The use of decision matrices, dominance reasoning, and the MEU Principle is best studied through examples. A good starting point is Alan Hájek's Stanford Encyclopedia entry on Pascal's Wager (2017), which carefully dissects exercise 1.10.

General rules for how to identify the acts, states, and outcomes for a decision problem can be found in chapter 2 of James Joyce's *The Foundations of Causal Decision Theory* (1999). The details are hard.

You may have come across an alternative definition of expected utility, using conditional probabilities and without a requirement that states be independent of the acts. We'll look at this formulation in chapter 9.

The maximin rule belongs to a family of decision rules that don't take into account the probability of the states. Such rules are sometimes thought to be needed for "decision-making under uncertainty", where – unlike in cases of "decision-making under risk" – the agent lacks information about the relevant probabilities. This makes sense if we assume (as many authors do) that the probabilities that figure in the definition of expected utility are objective quantities. In our Bayesian model, the probabilities are simply degrees of belief, and there is no such thing as "decision-making under uncertainty", where probabilistic information is unavailable. One advantage of the Bayesian approach is that it is hard to find a sensible decision rule that doesn't involve probabilities. Even imprecise probabilities – the topic of the essay question – raise serious problems: see Adam Elga, "Subjective Probabilities Should Be Sharp" (2010).

For a quick introduction to the problem of intentionality and the possibility of a decision-theoretic answer, see Ansgar Beckermann, "Is there a problem about intentionality?" (1996).

For some background on scientific modelling and idealisations, see Alisa Bokulich, "How scientific models can explain" (2011), and Mark Colyvan, "Idealisations in normative models" (2013).

Harsanyi's argument about utilitarian voters appears in his 1980 paper "Rule utilitarianism, rights, obligations and the theory of rational behavior". For more on the expected good caused by voting, not flying, and the like, see chapter 6 of William MacAskill, *Doing Good Better: Effective Altruism and a Radical New Way to Make a Difference* (2015).

The Miners Problem is from Nico Kolodny and John MacFarlane, "Ifs and Oughts" (2010).

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 in 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. My credence that the 37 bus goes to Waverley, for example, might be around 0.8, while my credence 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 give 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 this usage, the probability of an outcome is the proportion of that type of outcome in some base class of events. For example, on the textbook definition, to say that the probability of getting a six when throwing a regular die is $1/6$ is to say 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 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 I talk about credence or subjective probability, I do not mean belief about objective probability. I simply mean degree of belief. Our Bayesian model here diverges from *frequentist* or *objectivist* models that define expected utility in terms of objective probability. The MEU Principle is then restricted to cases in which the agent knows the relevant objective probabilities. (I mentioned this under "Sources and Further Reading" in the previous chapter.) On the Bayesian conception of probability, the MEU Principle does not presuppose knowledge of probabilities; it only presupposes that the agent 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, a structure that is studied by the mathematical discipline of probability theory.

Mathematically, a **probability measure** is a certain kind of function – in the mathematical sense: 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 (possibly 0), 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’).

(Here and henceforth, I use upper-case letters A, B, C, etc. as schematic variables for arbitrary propositions.)

In our application, Booleanism implies that if an 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 these in terms of negation, conjunction, and disjunction.

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 haven't defined ‘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. Consider the current temperature in Edinburgh. I don't know what it is. One possibility (one possible state of the world) is that it is 10°C. There is also a possibility that it is 50°F. How are these related? Since 10°C = 50°F, the second possibility is not an *alternative* to the first. It is the very same possibility, expressed with a different unit. The sentences ‘It is 10°C in Edinburgh’ and ‘It is 50°F in Edinburgh’ 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 is nothing but 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, perhaps as the set { 7°C, 8°C, 9°C, 10°C, 11°C, 12°C }. The unspecific possibility obtains just in case one of the more specific possibilities obtains. A maximally specific state is called a **possible world** (in philosophy, and an ‘outcome’ in many other disciplines). We will 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 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 is a *subset* of B if all objects in A are also in B. A is a *superset* of B if all objects in B are also in A.
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, in set theory terms, does it mean that a proposition A entails a proposition B?

Exercise 2.2 ††

Not all objects of probability are possible states of the world. Booleanism entails that at least one object of probability is *impossible*. Can you explain why?

Let's continue with the mathematics of probability. A probability measure, I said, 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 given by the Russian mathematician Andrej Kolmogorov in 1933.

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

I used have 'Cr' here for the probability measure, as we will be mostly interested in subjective probability or credence. '$\mathrm{Cr}(A)$' is 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 we're mostly dealing with statements that hold for all agents at all times, so we can 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 subjective probability of a disjunction equals the sum of the probability of 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 how. Let A be any proposition.

Then $A \vee \neg A$ is logically necessary. By axiom (ii),

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

Since A and $\neg A$ are logically incompatible, axiom (iii) tells us that

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

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 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 or worlds. 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. (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 be $9 \cdot 1/36 = 1/4$.

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.

Some thorny technical problems arise if there are infinitely many worlds. 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. If there are too many worlds, however, this turns out to be incompatible with the mathematical structure of the real numbers. The most one can safely assume is that the 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 $\mathrm{Cr}(A)$ cannot be greater than $\mathrm{Cr}(B)$. (You may use the rules we have already derived.)

2.4 Conditional probability

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. Yet, on the supposition that it will snow, I am not at all confident that the temperature will be above 4°C. My *unconditional credence* in temperatures above 4°C is high, but 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 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. 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 our world

is a world 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 in section 9. For now, let's focus on the first, indicative kind of supposition. I 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.

The slash '/' (some authors use '|') is not a connective. $\mathrm{Cr}(A/B)$ is not the agent's credence in a special proposition designated by 'A/B'. (Never write things like '$\mathrm{Cr}(A/B/C)$' or '$\mathrm{Cr}(A \wedge (B/C))$'. These have no meaning.)

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 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 $A \wedge B$ 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 '$\mathrm{Cr}(B/A)$' as shorthand for '$\mathrm{Cr}(A \wedge B)/\mathrm{Cr}(A)$'. 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) iff $\mathrm{Cr}(A/B) = \mathrm{Cr}(A)$. 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.

Unlike causal independence, probabilistic independence is a feature of beliefs. Two propositions can be independent for one agent and 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 often also independent in the agent's degrees of belief. You may want to ponder why that is the case.

Exercise 2.6 †

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

Exercise 2.7 ††

Using the Ratio Formula and the Equivalence Rule, show that if A is (probabilistically) independent of B, then B is independent of A (assuming that $\mathrm{Cr}(A)$ and $\mathrm{Cr}(B)$ are greater than 0).

Exercise 2.8 ††

A fair die will be tossed, and you give equal credence to all six outcomes. Let *Ex* be the proposition that the die lands 1 or 6. Let *Odd* be the proposition that the die lands an odd number (1, 3, or 5), and let *Low* be the proposition that the die lands 1, 2 or 3. Which of the following are true, in your belief state?

(a) *Ex* is independent of *Odd*.
(b) *Odd* is independent of *Ex*.
(c) *Ex* is independent of *Low*.
(d) *Odd* is independent of *Low*.

(e) *Ex* is independent of $Odd \wedge Low$.

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, you can figure out that if A and B are true and C is false, then $A \wedge \neg(B \vee \neg(C \vee A))$ is false. Now suppose instead of the truth-value of A, B, and C, I give you their probability. Could you compute the probability of $A \wedge \neg(B \vee \neg(C \vee A))$? 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 $\mathrm{Cr}(B/A) = \mathrm{Cr}(B)$. 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. The result is a half of a half. And half

of a half is $1/2$ times $1/2$.

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

The Disjunction Rule

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

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

Why do we subtract $\mathrm{Cr}(A \wedge B)$ in the Disjunction Rule? 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). By Additivity, the probability of the disjunction $A \vee B$ equals 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 $\mathrm{Cr}(A)$ and $\mathrm{Cr}(B)$, we have effectively double-counted the $A \wedge B$ worlds. That's why we need to subtract $\mathrm{Cr}(A \wedge B)$.

Exercise 2.9 †

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

Exercise 2.10 ††

Prove from the Ratio Formula that $\mathrm{Cr}(A \wedge B \wedge C) = \mathrm{Cr}(A/B \wedge C) \cdot \mathrm{Cr}(B/C) \cdot \mathrm{Cr}(C)$. (This is known as the *Chain Rule*, and generalizes to more than three conjuncts.)

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 that the chance of SIDS affecting both sons was 1/8500 · 1/8500 = 1 in 73 million. What is wrong with Sir Meadow's reasoning?

I want to mention two more rules that play a special role in Bayesian accounts. 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 probability of A given B to the inverse conditional probability of 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 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 $\mathrm{Cr}(\mathit{Six} \;/\; \mathit{Trick}) = 1$ and that $\mathrm{Cr}(\mathit{Trick}) = 1/2$. Moreover, $\mathrm{Cr}(\mathit{Six} \;/\; \neg\mathit{Trick})$ is plausibly $1/6$ and $\mathrm{Cr}(\neg\mathit{Trick})$ is $1/2$. Plugging all these values into the formula, we get $\mathrm{Cr}(\mathit{Trick} \;/\; \mathit{Six}) = 6/7$. 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 there are only two sexes, which are equally common and independent among siblings.)

Essay Question 2.1

If an agent's degrees of belief satisfy the probability axioms, it seems to follow from Kolmogorov's axiom (ii) that the agent must be certain of every logical truth. Does this mean that our Bayesian model is inapplicable to ordinary agents, who are not logically omniscient? If so, is this a problem? Do you have an idea of how the model could be adjusted to allow for logical non-omniscience?

Sources and Further Reading

There are many good introductions to elementary probability theory. For a slightly more in-depth discussion of the topics we have covered, you may want to consult chapters 3–7 of Ian Hacking, *An Introduction to Probability and Inductive Logic* (2001). (You may find the rest of the book helpful as well.)

The problems infinitely many worlds raise for the Additivity axiom are nicely explained in Brian Skyrms, "Zeno's paradox of measure" (1983).

The topic of the essay question is commonly discussed as the "problem of logical omniscience". See, for example, Zeynep Soysal, "A metalinguistic and computational approach to the problem of mathematical omniscience" (2022) for an interesting recent proposal with pointers to the earlier discussion.

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 rational degrees of belief satisfy the Booleanism condition from p.24. The remaining question is whether they should satisfy 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.

A natural thought is 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 agreed to put the upper limit at 1, it doesn't make 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 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 special feeling of conviction people have when entertaining propositions. On this approach, axiom (ii) says that when agents entertain logically necessary propositions, they ought to experience the relevant 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. In short, if we understand degrees of belief as measuring the intensity of a certain feeling, then the norms of Bayesian decision theory and Bayesian epistemology become implausible and inexplicable.

A more promising line of thought assumes that strength of belief is defined, perhaps in part, by the MEU Principle. On this approach, what we mean when we say that an agent has such-and-such degrees of belief is (in part) 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.

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

3.2 The betting interpretation

It is instructive to compare degrees of belief with numerical quantities in other parts of 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', clarifies which convention we're using.)

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 the same 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 object with greater mass will go down. We could now 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 – the average of the possible payoffs, weighted by their subjective probability. Let x be your degree of belief that it will rain tomorrow, and $1-x$ your degree of belief that it won't rain. The bet gives you £1 with probability x and £0 with probability $1-x$. The expected payoff is $x \cdot £1 + (1-x) \cdot £0 = £x$. This suggests that the bet is worth £x, that £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 is raining and £0 otherwise. 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, where a **unit bet on** a proposition A is a deal that pays £1 if A is true and £0 otherwise.

The betting interpretation

An agent believes a proposition A to degree x just in case she would buy a unit bet on A for up to £x (and she would sell a unit bet for A for £x or more).

Selling a bet means offering it to somebody else, in exchange for a fixed amount of money.

Exercise 3.2 ††

Show that selling a unit bet on A for £x is equivalent to buying a unit bet on $\neg A$ for £$(1-x)$, in the sense that the two transactions have the same net effect on the decision-maker's wealth, whether or not A is true.

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 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 she is disposed to enter bets that amount to a guaranteed loss.

3.3 The Dutch book theorem

In betting jargon, a combination of bets that are bought or sold is called a 'book'. A book that amounts to a guaranteed loss is called a '**Dutch book**' (no-one knows why). We are going to show 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 will be prepared to accept a Dutch book.

We begin with Kolmogorov's axiom (i). Suppose an agent's credence in some proposition A is greater than 1. Let's say it is 2. By the betting interpretation, the agent is willing to pay up to £2 for a deal that pays her back either £0 or £1, depending on whether A is true. She is guaranteed to lose at least £1. More generally, if an agent's degree of belief in A is greater than 1, then she will be prepared to buy a unit bet on A for more than £1, which leads to a guaranteed loss.

Similarly, suppose an agent's credence in A is below 0. Let's say it is -1. The agent will then be prepared to sell a unit bet on A 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 the agent would pay up to £1 for us to take the bet. Having sold the bet, she will have to pay us an additional £1 if A is true. Her net loss is either £2 or £1, and guaranteed be 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 ††

Show that if an agent's degrees of belief violate Kolmogorov's axiom (ii) then (assuming the betting interpretation) they are prepared to buy or sell bets that amount to a guaranteed loss.

Turning to axiom (iii), suppose an agent's credence in the disjunction $A \vee B$ of two logically incompatible propositions A and B is not the sum of her credence in the individual propositions. For concreteness, suppose $\mathrm{Cr}(A) = 0.4$, $\mathrm{Cr}(B) = 0.2$, and $\mathrm{Cr}(A \vee B) = 0.5$. By the betting interpretation, the agent is willing to sell a unit bet on $A \vee B$ for at least £0.50. She is also willing to buy a unit bet on A for up to £0.40, and she is willing to buy a unit bet on B for up to £0.20. Notice that if she buys both of these latter bets then she has in effect bought a unit bet on $A \vee B$, for she will get £1 if either A or B is true, and £0 otherwise. So the agent is, in effect,

willing to buy this bet for £0.60 and sell it for £0.50. You can check that no matter whether A or B or neither of them is true, the agent is guaranteed to lose £0.10.

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

We have proved the *Dutch Book Theorem*.

Dutch Book Theorem

Assuming the betting interpretation, any agent whose degrees of belief don't conform to the Kolmogorov axioms is prepared to buy bets whose net effect is a guaranteed loss.

One can also show the converse, that any agent who is prepared to accept a (certain kind of) Dutch book has non-probabilistic beliefs. In other words, agents whose beliefs conform to the rules of probability are *not* prepared to accept (certain kinds of) bets that amount to a guaranteed loss. This result is known as a **Converse Dutch Book Theorem**. I'll outline a proof.

To get an interesting Converse Dutch Book result, we should extend the betting interpretation so that it doesn't just cover unit bets. (We don't just want to show that an agent with probabilistic beliefs is not prepared to accept a Dutch Book *made entirely of unit bets*.) Let's assume that agents generally value bets by their expected monetary payoff, so that they pay up to £x for a bet with expected payoff £x, where the expected payoff is computed with the agent's credence function. We're now interested in cases where this credence function is a genuine probability measure, so that the expected payoff is a genuine "expectation", in the mathematical sense: a probability-weighted average.

Now consider an agent with probabilistic beliefs. If the agent pays some amount £x for a bet with expected payoff £y, then the entire transaction (including the purchase price) has expected payoff £$(y–x)$. By our extended betting interpretation, the agent makes the transaction only if $x \leq y$, in which case £$(y–x) \geq$ £0. In other words, the agent makes the transaction only if the transaction has a non-negative expected payoff. Evidently, a transaction can't have a non-negative *expected* payoff unless there is at least some possibility for it to have a non-negative payoff. This shows that an agent with probabilist credences can't be "Dutch booked" with a single bet.

What about combinations of bets? Suppose our agent buys a number of bets. We know that each of these transactions on its own has a non-negative expected payoff. We also know that the total payoff from all transactions together is the sum of the payoffs of the individual transactions. Now here is a useful fact about mathematical expectation: *the expectation of a sum (of some quantities) is the sum of the expectations (of the quantities)*. Since the sum of non-negative values can't be negative, this tells us that the expected total payoff from our agent's transactions isn't negative. As before, we can infer that the combined transactions are not guaranteed to generate a loss.

Exercise 3.4 ††

Here I have twice appealed to the fact that if a transaction or combination of transactions has non-negative *expected* payoff, then there must be at least a possibility of an *actual* non-negative payoff. Can you explain why this is the case? Does it depend on whether the expected payoff is computed with a genuine probability function?

Exercise 3.5 ††

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.

A flat-footed "Dutch book argument" might go as follows. If your beliefs violate the probability axioms, then a cunning Dutchman might come along and trick you out of money. If your beliefs are probabilistic, he can't do that. To be safe against the Dutchman, it is better to have probabilistic beliefs.

Is this a good argument for probabilism? Two problems stand out.

First, why should the 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. If there is such a Frenchman, we can at most conclude that it would be *practically useful* to have non-probabilistic beliefs. But those beliefs would still be *epistemically irrational*. (Compare: if someone offers you a million pounds if you believe that the moon is made of cheese, then that belief would be practically useful, but it would not be epistemically rational.) 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, as implied by the betting interpretation. Real people don't actually do this.

Consider the following gamble.

Example 3.1 (The St. Petersburg Paradox)

I am going to toss a fair coin until it lands tails. If I get tails on the first toss, I'll give you £2. If I get heads on the first toss and tails on the second, I'll give you £4. If I get heads on the first two tosses and tails on the third, I'll give you £8. In general, if the coin first lands tails on the nth toss, I'll give you £2^n.

How much would you pay for this gamble?

We can compute the expected payoff. With probability ½ you'll get £2; with probability ¼ you get £4; with probability ⅛ you get £8; and so on. The expected payoff is

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

The sum of this series is infinite. 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 3.6 †

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

The St. Petersburg Paradox was first described by the Swiss mathematician Nicolas Bernoulli in 1713. It prompted 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 measures how much the agent desires a small extra amount of the good. That the marginal utility of money is declining means that the more money you have, the less you value an additional pound (or dollar or ducat).

Bernoulli had a more concrete proposal. He suggested that n units of money provide not n but $\log(n)$ units of utility. This implies that doubling your wealth always provides the same boost in utility, whether it leads from £1000 to £2000 or 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 equivalent to the utility of getting £10.95. That's the most the agent should be willing to pay for the gamble.

Exercise 3.7 †

Suppose Bernoulli is right that owning £n has 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 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 3.8 ††

As Bernoulli noticed, the declining marginal utility of money can explain the usefulness of insurance. Suppose your net worth is £10 000, and there's a 5% chance of a catastrophic event that would cost you £9 000. For a fee of £1 000, a bank offers you an insurance against the catastrophic event that pays £9 000 if the event occurs (and nothing otherwise). Explain (informally, if you want) why this might be a good deal both for you and for the bank.

Exercise 3.9 †

Bernoulli's logarithmic model is obviously a simplification. Suppose you want to take a bus home. The fare is £1.70 but you only have £1.50. 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 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 offer.

There's a second reason why rational agents wouldn't always value bets by their expected payoff even if their subjective utility were 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 unit bet 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, but it would be irrational for me to buy it even for £0.10.

In sum, we can't assume that rational agents always value bets by their expected payoff. The betting interpretation is indefensible.

This is a setback on two fronts. One, we have lost an attractive answer to how degrees of belief are measured or defined. If an agent's degrees of belief aren't defined by their betting behaviour, then how *are* they defined? Second, and relatedly, we have lost what looked like an attractive argument for probabilism. If agents don't value bets by their monetary payoff, we can't show that non-probabilistic agents will be prepared to buy bets that amount to a sure loss.

We will look at alternative approaches to measuring belief in sections 3.6 and 6.5.

First, let me explain how we might rescue an argument for probabilism from the wreckage of the betting interpretation.

3.5 A Dutch book argument

We want to show that non-probabilistic beliefs are irrational. Let α be an arbitrary agent with non-probabilistic beliefs. We can't assume that α values bets by their expected monetary payoff. But let's imagine a counterpart β of α who has the exact same beliefs as α but possibly different, and somewhat peculiar desires. β's only goal is to increase her wealth. Money does not have declining marginal utility for β. She would give all she has for an opportunity to play the St. Petersburg gamble. β might also differ from α in another respect: whenever she faces a choice, β chooses an option that maximizes expected utility.

I'm going to need a number of philosophical assumptions. Here is the first: *if α's belief state is epistemically rational, then so is β's*. The idea is that if you want to know if someone's beliefs are epistemically rational (rather than, say, practically useful), then you need to know what her beliefs are and maybe how she acquired those beliefs, but you don't need to know what she desires or how she chooses between available acts.

As we saw at the end of the previous section, we can't assume that β will always pay up to £Cr(A) for a unit bet on A (where Cr is her credence function), 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 she buys or sells any particular bets. 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. In particular, there aren't special consistency requirements that only pertain to stable beliefs. *If an agent's credences over stable propositions should be probabilistic, then her entire credence function should be probabilistic.* This is my second assumption. It implies that in order to show that non-probabilistic beliefs are irrational, it suffices to show that non-probabilistic beliefs towards 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.

We know that if a proposition A is stable, then β is prepared to pay up to £Cr(A) for a unit bet on A. That's because β's utility function simply measures monetary payoff and because she obeys the MEU Principle. The betting interpretation is correct for β, as long as we stick with stable propositions.

We also know that β's credences towards stable propositions violate the probability axioms. It follows by the Dutch Book Theorem that she is prepared to buy bets whose net effect is a guaranteed loss. My next assumption states that it would be irrational for β to make these transactions: *it is irrational for an agent whose sole aim is to increase her wealth to (deliberately and avoidably) make choices whose net effect is a guaranteed loss*.

This was my third assumption. My fourth assumption is that irrational choices always arise from either irrational beliefs or from irrational desires or from an irrational way of linking up one's beliefs and desires to one's actions. I also assume that the right way of linking up beliefs and desires to actions is given by the MEU Principle. Thus: *if an agent is disposed to make irrational choices, then she is either 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. Her choices do maximize expected utility. I also claim (assumption 5) that *β's desires are not irrational*. Admittedly, her desires are odd. We might call them unreasonable, or even "irrational" in a substantive sense. But they aren't inconsistent. They represent a coherent evaluative perspective.

Since β is disposed to make irrational choices, we can infer that she is epistemically irrational. By the very first assumption, it follows that α is epistemically irrational. And α was an arbitrary agent whose credences violate the rules of probability. We've shown that (epistemically) rational beliefs are probabilistic.

My argument relies on a lot of assumptions. Many of them could be challenged. Can you think of a better argument?

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. We need to explain how this convention of assigning numbers to propositions works.

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 (who, incidentally, also published the first proof of the Dutch Book Theorem). De Finetti suggested that degrees of belief might 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 we have $\mathit{Sunny} \succ \mathit{Rainy}$. 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 $A \succ B$ or $A \sim B$.

We now make some assumptions about the formal structure of these relations. To begin, if you are more confident in A than in B, then you can't also be more confident in B than in A, or equally confident in the two. We also assume that if you're neither more confident in A than in B, nor in B than in A, then you're equally confident in A and B. Your comparative credence relations are then "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 more confident in A than in B, and more confident in B than in C. Then you should be more confident in A than in C. Similarly, if you are equally confident in A and B, and in B and C, then you should be equally confident in A and C. So $\succ$ and $\sim$ should be "transitive":

Transitivity

If $A \succ B$ and $B \succ C$ then $A \succ C$; if $A \sim B$ and $B \sim C$ then $A \sim C$.

Exercise 3.10 †††

Suppose we define $A \sim B$ as $\neg(A \succ B) \wedge \neg(B \succ A)$. Show that Completeness is then entailed by the assumption that if $A \succ B$ then $\neg(B \succ A)$.

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).

Non-Trviality

$\top \succ \bot$.

Boundedness

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

These should be fairly plausible demands of rationality.

My next assumption is best introduced 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 $A \succsim B$ iff $\mathrm{Cr}(A) \geq \mathrm{Cr}(B)$ (which entails that $A \succ B$ iff $\mathrm{Cr}(A) > \mathrm{Cr}(B)$ and $A \sim B$ iff $\mathrm{Cr}(A) = \mathrm{Cr}(B)$). 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, Non-Triviality, Boundedness, Quasi-Additivity, and the Sixth Assumption, then there is a unique probability measure Cr such that $A \succsim B$ iff $\mathrm{Cr}(A) \geq \mathrm{Cr}(B)$.

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 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 one proposition than in another, the first goes to the right of the first. Whenever the agent is equally confident in two propositions, they are stacked on top of each other at the same point on the line. If the agent is reasonable, the impossible proposition $\bot$ will be at the left end, the necessary proposition $\top$ at the right end.

We now want to use numbers to represent the relative position of propositions along the line, in such a way that as we move from the $\bot$ position to the $\top$ position, the numbers get higher and higher. The Probability Representation Theorem assures us that this can be done, provided that the agent's comparative judgements satisfy the six assumptions. In that case, it says, there will be an assignment of numbers to propositions that "represents" the agent's comparative judgements in the sense that $A \succsim B$ iff the number assigned to A is at least as great as the number assigned to B.

The next problem is that if there is one such assignment then there are infinitely many, giving different numbers to propositions in between $\bot$ and $\top$. (For example, if f represents $\succsim$ then so does the function g defined by $g(A) = f(A)^2$.) We need to settle on a particular assignment. Again, the Probability Representation Theorem comes to our help. It tells us that among the eligible assignments of numbers to propositions – among those that represent the agent's comparative judgements – there is only one that satisfies the conditions on a probability measure. Let's adopt the convention of using this assignment.

On this approach, '$\mathrm{Cr}(\mathit{Rain}) = 0.2$' means that the agent's comparative confi-

dence judgements order the propositions in such a way that the unique probability measure that "represents" these judgements assigns 0.2 to *Rain*. 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 (!) that represents her comparative judgements. An agent who doesn't satisfy the six assumptions doesn't have a credence function at all, because our convention of measurement – on the present approach – doesn't cover such agents.

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 constraints on any agent with degrees of belief. The missing sixth assumption is especially troublesome in this regard. The form of the assumption turns out to depend on whether the number of propositions is finite or infinite. In either case the condition is so complicated that many struggle to accept it as a basic norm of rationality – let alone as a basic condition anyone must satisfy in order to have degrees of belief at all. 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 there are equally many propositions in the first sequence that contain w as in the second, if $A_i \gtrsim B_i$ for all $i < n$, then $B_n \gtrsim A_n$.

Essay Question 3.1

I have expressed the Dutch Book Theorem with monetary outcomes. One might try to avoid commitment to the betting interpretation by replacing the monetary outcomes with other goods the agent happens to care about. For example, when we looked at Kolmogorov's axiom (i), I said that an agent whose degree of belief in A is 2 would pay (say) £1.50 for a bet that pays £1 if A is true and £0 otherwise. This assumes the betting interpretation. Now let 'U1.5' denote an arbitrary good to which the agent assigns utility 1.5. Similarly, let

U1 be a good with utility 1, and U0 a good with utility 0. Consider a bet that would give the agent U1 if A is true and U0 otherwise. The bet's expected utility is $\mathrm{Cr}(A) \cdot \mathrm{U}(\mathrm{U1}) + \mathrm{Cr}(\neg A) \cdot \mathrm{U}(\mathrm{U0}) = \mathrm{Cr}(A)$. Assuming the MEU Principle, an agent with $\mathrm{Cr}(A) = 2$ would prefer this bet over U1.5, even though the latter is guaranteed to give her greater utility, which is surely irrational. Can you spell out a full argument for probabilism along these lines? What problems do you see for this line of argument?

Sources and Further Reading

For a critical overview and assessment of Dutch book arguments, see Alan Hájek, "Dutch Book Arguments" (2008). If you want to dive even deeper, you may start with Susan Vineberg's Stanford Encyclopedia entry on Dutch Book Arguments (2022).

For a more extensive philosophical introduction and criticism of the comparative approach from section 3.6, see Edward Elliott, "Comparativism and the Measurement of Partial Belief" (2020). Peter Fishburn's "The Axioms of Subjective Probability" (1986) goes deeper into the mathematical background.

A recently popular third way of arguing for probabilism, besides the Dutch book approach and the comparative approach, draws on the observation (also first made by de Finetti) that for every non-probabilistic credence function there is a probabilistic credence function that is guaranteed to be closer to the truth – where closeness to the truth is a certain measure of the distance between the credence given to any proposition and the proposition's truth-value (0=false, 1=true). See, for example, James Joyce, "A nonpragmatic vindication of probabilism" (1998).

Martin Peterson's Stanford Encyclopedia entry on the St. Petersburg paradox discusses the historical context of the St. Petersburg paradox and also introduces a "modern" version in which the monetary payoffs are replaced by units of utility.

The bus fare exercise is from Brian Skyrms, *Choice and Chance* (2000).

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 to her desires and choices. The second, probabilism, imposes an internal, structural constraint on rational beliefs: that they conform to the rules of probability. 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.

Seeing the red paper should increase your confidence that the liquid is acidic. 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 norms on how beliefs change in response to perceptual experience. Like the MEU Principle, and unlike probabilism, these norms state a connection between beliefs and something other than belief – here, perceptual experience. Loosely speaking, the MEU Principle describes the causal "output" of beliefs: the effects an agent's beliefs have on her behaviour. Now we turn to the "input" side. We want to know what sorts of experiences might cause a rational agent to have such-and-such beliefs.

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

We could try to identify the 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.

We could alternatively identify perceptions 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 and remain neutral on details of biological or electrical implementation.

The usual strategy is to identify perceptual experiences by the information they provide to the agent's belief system. In the Litmus Test, for example, we might assume that the information you receive from your visual system is that the paper has turned red. You don't directly receive the information 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 their perceptual experiences can always be captured by a single 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

Suppose a perceptual experience provides an agent with some information E (for "evidence"). How should the rest of the agent's 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 when you see 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 your liquid is acidic *on the supposition that* the paper will turn red. Your initial degrees of belief might have been as follows.

$\mathrm{Cr}_{\mathrm{old}}(\mathit{Acid}) = 1/2$.
$\mathrm{Cr}_{\mathrm{old}}(\mathit{Acid} \mid \mathit{Red}) = 9/10$.

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 conditionalize **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:

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

Here it is understood that the agent's 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. If you see the paper turn red but at the same time notice a whiff of ammonium hydroxide, which you know is alkaline, your credence in the *Acid* hypothesis may not increase to 0.9.

Exercise 4.1 †

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

Exercise 4.2 ††

Show that conditionalizing first on E_1 and then on E_2 is equivalent to conditionalizing in one step on $E_1 \wedge E_2$. That is, if Cr_1 results from Cr_0 by conditionalising on E_1, and Cr_2 results from Cr_1 by conditionalizing on E_2, then for any A, $\mathrm{Cr}_2(A) = \mathrm{Cr}_0(A \mid E_1 \wedge E_2)$. (You may assume that $\mathrm{Cr}_0(E_1 \wedge E_2) > 0$.)

Exercise 4.3 †††

Assume that $\mathrm{Cr}_{\mathrm{new}}$ results from $\mathrm{Cr}_{\mathrm{old}}$ by conditionalizing on some information E with $\mathrm{Cr}_{\mathrm{old}}(E) > 0$, and that $\mathrm{Cr}_{\mathrm{old}}$ satisfies the Kolmogorov axioms. Using the probability rules, show that $\mathrm{Cr}_{\mathrm{new}}$ also satisfies the Kolmogorov axioms. (You may use any of the derived rules from chapter 2. Hint for axiom (ii): if A is logically necessary, then $A \wedge E$ is logically equivalent to E.)

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**:

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

This formulation is useful because it is often easier to evaluate $\mathrm{Cr}_{\mathrm{old}}(E/A)$, the probability of the evidence E conditional on some hypothesis A, than to evaluate $\mathrm{Cr}_{\mathrm{old}}(A/E)$, the probaility of the hypothesis conditional on the evidence.

Here is 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 positive result in 10% of non-cancer cases. A woman from the population comes into your practice, takes the test, and gets a positive result. How confident should you be that the woman has breast cancer?

We assume that you are aware of all the statistical facts before you learn 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, call it C, that the woman has cancer should plausibly be 0.02. So we have

$$\mathrm{Cr}_{\mathrm{old}}(C) = 0.02.$$

Since you know that the test yields a positive result in 95% of cancer cases, we also

have

$$\mathrm{Cr}_{\mathrm{old}}(P/C) = 0.95,$$

where P says that the test result is positive. Similarly, since the test yields a positive result in 10% of non-cancer cases, we have

$$\mathrm{Cr}_{\mathrm{old}}(P/\neg C) = 0.1.$$

Now we simply plug these numbers into Bayes' Rule, expanding the denominator by the Law of Total Probability:

$$\begin{aligned}\mathrm{Cr}_{\mathrm{new}}(C) &= \frac{\mathrm{Cr}_{\mathrm{old}}(P/C) \cdot \mathrm{Cr}_{\mathrm{old}}(C)}{\mathrm{Cr}_{\mathrm{old}}(P/C) \cdot \mathrm{Cr}_{\mathrm{old}}(C) + \mathrm{Cr}_{\mathrm{old}}(P/\neg C) \cdot \mathrm{Cr}_{\mathrm{old}}(\neg C)} \\ &= \frac{0.95 \cdot 0.02}{0.95 \cdot 0.02 + 0.1 \cdot 0.98} = \frac{0.019}{0.019 + 0.098} = 0.16.\end{aligned}$$

After the positive test, your degree of belief that the woman has breast cancer should be 0.16. This is lower than many people initially think – including many trained physicians. But it makes sense. Imagine we took a 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% = 98 to test positive. The total number of positive tests would be around 19 + 98 = 117. Of these 117 women, 19 actually have cancer. So the chance that an arbitrary 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 resembles this statistical line of reasoning.

The tendency to overestimate (or underestimate) probabilities 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 have chosen box A?

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

A murder has been committed on an 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. 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 this true? As a member of the jury, how confident should you be in the defendant's guilt?

4.3 Induction and Indifference

Suppose an agent's beliefs are probabilistic and change by conditionalization. Does this ensure that the beliefs are reasonable? No. If the agent starts out with sufficiently crazy beliefs, conditionalization will not make them sane.

Example 4.3

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

You should be fairly confident that the next bird will also be green. The Principle of Conditionalization does not ensure this. It might even make you confident that the next bird is pink. For suppose you were born with a firm conviction that if you are ever going to see 100 green birds on an island, then the next bird you would see is pink. Your observation of 100 green birds does not challenge this conviction. After conditionalizing on your observation of the 100 green birds, you would become confident that the next bird you will encounter is pink.

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 probably green we need a further premise about the "uniformity of nature". Roughly, we need to assume

that regularities in the part of the world that we have observed up to some time are likely to extend into the unobserved part of the world. If, for example, the first 100 birds we encounter on an island are all green, then other birds on the island are probably green as well. This assumption may be supported by earlier experiences. But, again, it won't be *entailed* by these experiences. Ultimately, some such premise must be accepted as bedrock.

In Bayesian terms, the problem of induction suggests that we have to put restrictions on what an agent may believe *without any relevant evidence*. Scientifically minded people sometimes feel uneasy about such restrictions, and therefore speak about the **problem of the priors**. An agent's **priors** (or "ultimate priors" or "urpriors") are her credences at the start of her epistemic journey, before she conditionalizes on any evidence.

What should an agent believe, at the beginning of her epistemic journey? It would be irrational to be convinced, without any evidence, that the first 100 birds one might encounter on an island will be atypical in colour. Indeed, a natural thought is that without any relevant evidence, one should not be convinced of anything (except logical truths). One should be open-minded, dividing one's credence evenly between all ways the world might be:

The (naive) Principle of Indifference

If $A_1, \ldots, A_n$ are n propositions exactly one of which must be true, then a rational prior credence function assigns the same probability $1/n$ to each of these propositions.

This, however, can't be right. 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 naive Principle of Indifference, you should give credence $1/2$ to R and $1/2$ to $\neg R$. 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.

By the naive Principle of Indifference, you should give credence $1/5$ to each of these possibilities. 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 *multiple* 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 briefly return to it in section 11.5.

A more plausible response is to restrict the propositions $A_1, \ldots, A_n$ to which the requirement of indifference applies. Intuitively, you shouldn't be indifferent between R and $\neg R$ because these two propositions are not on a par. There are more ways of being non-red than of being red. Unfortunately, it is hard to turn this intuition into a consistent general rule, as the following exercise illustrates.

Exercise 4.6 ††

I have a wooden cube in my office whose side length is at least 2 cm and at most 4 cm. That's all you know about the cube. We can distinguish two possibilities:

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

The intervals have the same length, so S and L are intuitively on a par. This suggests that you should give credence $1/2$ to each of 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 credence do you give to S and L if you treat equally sized ranges of volume as equally likely?

There is another problem with indifference principles. Let's imagine we've found a rule for when two propositions are "on a par" so that we can consistently require

an agent's priors to be indifferent between propositions that are on a par. We should still be cautious about endorsing the requirement, for is likely to clash with the "uniformity of nature" assumption required for inductive inference.

Return to example 4.3. Assume, for simplicity, that birds can only be green or red. There are then four possibilities regarding the first two birds you might see:

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

Intuitively, these four possibilities are on a par. An indifference principle might say that you should give credence $1/4$ to each.

Now what happens when you see the first bird, which is green? Your evidence rules out *RG* and *RR*. If you conditionalize on your evidence, your new credence will be divided evenly between the remaining possibilities *GG* and *GR* (as you may check). Your credence that the next bird is green will be $1/2$. By the same reasoning, if your prior credence is evenly divided between all possible colour distributions among the first three birds (*GGG*, *GGR*, etc.), then after having seen two green birds, your "posterior" credence that the next (fourth) bird is green will still be $1/2$. And so on. No matter how many green birds you see, you won't think that this tells you anything about the next bird.

If we want an observation of 100 green birds to raise your credence in the next bird being green, we have to assume that your prior credence in the "uniform" hypothesis $GG\ldots GG$ (that's 101 '*G*'s) should be greater than your prior credence in the "non-uniform" hypothesis $GG\ldots GGR$.

Intuitively, the problem is that there are at least as many irregular worlds as regular worlds. If you spread your credence evenly over all ways the world might be, you'll end up giving too much credence to irregular worlds. You won't be able to learn by induction.

Rational priors should be open-minded, but biased towards regular worlds. There is no agreement on how to make this precise. (We will meet an intriguing partial answer in the following section.) As such, the "problem of the priors" remains open.

4.4 Probability coordination

We turn from the highly controversial Principle of Indifference to another norm that is almost universally accepted among Bayesians. This norm connects subjective probability with objective probability, and is often expressed as a norm on priors.

The Probability Coordination Principle

An agent's prior credence in a proposition A, on the supposition that the objective probability of A is x, should equal x:

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

Here, Cr_0 is a rational prior credence function, and Pr is any kind of objective probability, such as relative frequency or quantum physical chance.

The Probability Coordination Principle implies that if a rational agent has discovered the objective probabilities – if she has conditionalized on $\Pr(A) = x$ – and she doesn't have other relevant information about A, then she will align her degrees of belief with the objective probabilities: her degree of belief in A will match the known objective probability.

We have unwittingly assumed this all along. In example 4.2, we assumed that if you know that a woman is from a population in which 2% of women have cancer, and you have no other relevant information about her, then your credence that she has cancer should be 0.02. This is not entailed by the Kolmogorov axioms. We need the Probability Coordination Principle to connect your information about relative frequency to your degree of belief.

The Probability Coordination Principle can be used even if the agent doesn't have full information about the objective probabilities. In exercise 4.4, you had to evaluate $\mathrm{Cr}(\mathit{Black} \mid B)$, where B is the hypothesis that I have drawn a ball from a box containing one black ball and one white ball, and *Black* is the hypothesis that the ball is black. Assuming that the draw is random (in some objective sense), B entails that $\Pr(\mathit{Black}) = 1/2$. You don't know whether B is true, but we can infer, by the Probability Coordination Principle, that $\mathrm{Cr}(\mathit{Black} \mid B) = 1/2$.

In 1814, Pierre-Simon Laplace observed that the Probability Coordination Principle may help with Hume's problem of induction. Return to example 4.3, where

you've encountered 100 green birds in the first few days on a remote island. Suppose you think that there's a certain objective probability with which any given bird on the island is green (independently of the other birds). That probability might be 1, in which case all the birds are certain to be green. Or it might be 0. Or it might be anything in between 0 and 1, in which case you would expect to find some red birds and some green birds. Now suppose you start out maximally open-minded about this probability, giving equal credence to all values from 0 to 1. Using the Probability Coordination Principle, one can then show – the maths is beyond what we do in this course – that after observing 100 green birds, your credence that the next bird is green will be around 0.99. You have learned by induction!

In the previous section, we saw that indifference over *outcomes*, over possible sequences of G and R, makes inductive learning impossible. Laplace saw that indifference over (objective) *probabilities of outcomes* has the opposite effect. By treating the outcomes as independent matters of objective probability, and giving equal credence to the objective probability, you end up giving comparatively low credence to irregular sequences.

You may wonder where the Probability Coordination Principle comes from. Some say it is a basic norm of rationality. Others say that it must follow from more basic norms – from a restricted indifference principle, for example, or even from probabilism alone. The issue turns on deep questions about the nature of objective probability. Those who regard Probability Coordination as basic tend to believe that the ultimate fabric of the physical world includes probabilistic quantities to which rational beliefs should be aligned, for reasons nobody can explain. Those who don't regard Probability Coordination as basic see no need to posit special physical quantities with a mysterious spell on rational credence. On a simple version of the alternative view, objective probability is nothing but relative frequency and the Probability Coordination Principle follows from plausible indifference requirements. We will not look further into these debates.

Exercise 4.7 †

Jacob Bernoulli (an uncle of Daniel and Nicolas Bernoulli, who we've met in section 3.4) proposed the following simplified version of the Probability Coordination Principle: *If a proposition has very low objective probability, one may be certain that it is false*. What do you think of this?

4.5 Confirmation

An important question both in the philosophy of science and in scientific practice is how scientific hypotheses are confirmed or disconfirmed by empirical data. We can't directly observe that, say, spacetime is curved, that smoking causes cancer, or that dolphins evolved from land animals. Our evidence strongly *supports* these assumptions, but it doesn't entail them. What is this relation of evidential support? What does it take for some evidence to support a hypothesis?

Philosophers have tried to formulate general rules for evidential support, akin to the rules of deductive logic. The following rules, or "conditions" on when a hypothesis is confirmed by evidence, figure in an influential 1945 paper by Carl Hempel.

> **Nicod's Condition**. Universal generalisations are confirmed by their instances: an F that is G lends support to the hypothesis that all Fs are Gs.
>
> **Converse Consequence Condition**. If some evidence confirms a hypothesis then it also confirms any theory (conjunction of hypotheses) that entails the hypothesis.
>
> **Special Consequence Condition**. If some evidence confirms a theory then it also confirms anything that is entailed by the theory.

This rule-based (or "syntactical") approach didn't work out well. Most rules that initially looked plausible turned out to have clear counterexamples. The few that remained are too weak to make sense of scientific reasoning.

Consider Nicod's Condition. Normally, observation of a black raven lends support to the hypothesis that all ravens are black. But not always. Suppose your friend is on an expedition and you've agreed that if she comes across a white raven then she is going to send you a black raven, by mail, in a cage. One day, a parcel arrives: it's a black raven. In this context, observation of a black raven is strong evidence *against* the hypothesis that all ravens are black.

Exercise 4.8 ††

Show that the Converse Consequence Condition and the Special Consequence Condition together entail that if some evidence confirms some hypothesis then the same evidence confirms every hypothesis whatsoever.

A different kind of approach was suggested by Karl Popper. Popper noticed that although scientific theories are rarely entailed by empirical evidence, they can be *refuted* by the evidence. A single white raven is enough to refute (or "falsify") the hypothesis that all ravens are black. According to Popper, a theory is confirmed (or "corroborated", as he preferred to say) to the extent that it has withstood attempts at falsification.

One problem for this **falsificationist** approach is that many scientific theories or hypotheses can't actually be falsified, because they don't have directly observable consequences. The (well-confirmed) hypothesis that smoking causes cancer, for example, doesn't imply that every single smoker gets cancer. It only predicts that smokers have a higher *probability* of getting cancer, in some objective sense of 'probability'. (The hypothesis is not about what people believe.) We can't directly observe that probability.

To get around this issue, falsificationism may call upon its powerful ally, **"classical" (or "frequentist") statistics**. According to classical statistics, a hypothesis can be rejected not only if it is logically incompatible with the evidence, but also if it renders the evidence *sufficiently improbable*. Imagine, for example, that we randomly divide 1000 children into two groups. One group is instructed to take up smoking, the other to refrain from smoking. In all other respects, we force the two groups to lead similar lives. 50 years later, we find more incidents of cancer in the smoking group than in the "control group". This could be just a coincidence. The tools of classical statistics allow us to compute the objective probability of the observed difference between the groups *on the assumption that it is a coincidence*. If this probability is sufficiently low, classical statistics tells us that we can reject the coincidence hypothesis. We can infer that smoking really does increase the risk of cancer.

One obvious problem with this move is to explain when a probability is "sufficiently low". Just how improbable must a hypothesis render the observed evidence to warrant rejecting the hypothesis? In the social sciences, any probability below

0.05 is usually deemed sufficiently low. In medicine, a threshold of 0.01 is preferred. Either choice looks unprincipled and arbitrary. Besides, what does it mean to "reject" a hypothesis? Should we become absolutely certain that the hypothesis is false – even though we know that low-probability events happen all the time?

Another problem with the frequentist approach is that it is only applicable to specific kinds of data. The cancer experiment I have just described has never been carried out, for obvious ethical and practical reasons. The actual data that support the link between smoking and cancer are, for the most part, of a kind for which the tools of classical statistics aren't designed because one can't easily compute informative objective probabilities.

A deeper problem with the falsificationist/frequentist approach is that predictive success is not the only standard by which we evaluate scientific hypotheses. Physicists, for example, favour mathematically elegant theories, like Einstein's theory of General Relativity, that unify a diverse range of phenomena. Consider a rival hypothesis to Einstein's according to which the laws of General Relativity hold throughout all of space and time except tomorrow afternoon in my back garden, where nature obeys the laws of Aristotelian physics. This crackpot "theory" is logically compatible with all existing observations, and it doesn't render any of them less probable than Einstein's. By falsificationist lights, Einstein's theory and mine are equally well confirmed. Is that true? If you want to predict what is going to happen tomorrow afternoon in my back garden, you would surely be insane to rely on my theory.

A third approach to confirmation, besides the syntactical and the falsificationist approach, is **Bayesian Confirmation Theory**. It is by far the most popular approach in contemporary philosophy of science. (Its statistics ally is **Bayesian Statistics**.)

Why do we care about whether, or to what extent, a hypothesis is confirmed by the evidence? Ultimately, it's because we want to know how much credence we should invest in the hypothesis. We want to know how confident we should be that smoking causes cancer, or that the laws of Aristotelian physics will be operative tomorrow in my garden.

Bayesianism offers a simple, albeit schematic, answer. If E is the relevant evidence, then the credence we should give to a hypothesis H in light of E is $\mathrm{Cr}_0(H/E)$, where Cr_0 is a rational prior credence function.

In fact, Bayesians distinguish two notions of evidential support. We may ask about the *absolute* degree to which a hypothesis is supported by the evidence, but we may also ask about the *incremental* effect a single piece of data has on the credibility

of the hypothesis. One black raven, for example, hardly makes it probable that all ravens are black. Still, under normal circumstances, it lends some support to the generalisation.

The Bayesian analysis of confirmation

E **(absolutely) confirms** H to the extent that $\mathrm{Cr}_0(H/E)$ is high.
E **(incrementally) confirms** H to the extent that $\mathrm{Cr}_0(H/E)$ exceeds $\mathrm{Cr}_0(H)$.

Without more information about the prior credence Cr_0 these schematic analyses may not appear terribly useful. But let's have a closer look.

On the Bayesian account, confirmation comes in degrees, and its degree is closely related to the conditional probability $\mathrm{Cr}_0(H/E)$. With the help of Bayes' Theorem, we can break this conditional probability into three parts, which we may understand as three components of Bayesian confirmation:

$$\mathrm{Cr}_0(H/E) = \frac{\mathrm{Cr}_0(E/H) \cdot \mathrm{Cr}_0(H)}{\mathrm{Cr}_0(E)}.$$

The first component is $\mathrm{Cr}_0(E/H)$. This is the probability of the evidence given the hypothesis. The Bayesian analysis implies that, all else equal, the more probable the evidence is in light of a hypothesis, the more the evidence supports the hypothesis. Conversely, if a hypothesis renders the evidence unlikely, then (all else equal) the evidence is evidence against the hypothesis. In easy cases, we may use the tools of classical statistics to compute an objective probability for E given H, and invoke the Probability Coordination Principle to determine $\mathrm{Cr}_0(E/H)$. But we don't have to go via objective probabilities. We can take into account all kinds of data. And we don't need an arbitrary cutoff at which the hypothesis is "rejected".

The second component, $\mathrm{Cr}_0(H)$, is the prior probability of the hypothesis. This is where simplicity, systematicity, and other such criteria enter the picture. My crackpot theory about my Aristotelian back garden deserves negligible prior probability. (Why? Because rational priors assume that nature is "uniform", and my theory posits a bizarre kind of non-uniformity.)

The third component, $\mathrm{Cr}_0(E)$, is the prior probability of the evidence. This occurs in the denominator, meaning that the *lower* the prior probability of the evidence, the *higher* the degree of confirmation. This makes sense. Einstein's theory of Rel-

ativity predicts that light is deflected when it travels past massive objects. The first observation of this effect, in 1919, was deemed a great triumph for Einstein, because the observation was so surprising. It has low prior probability. By comparison, if an astrologer predicts that we will face personal challenges and make new acquaintances in the coming year, and the prediction comes true, this isn't a great triumph for astrology, because the prediction was highly probable all along.

So we can say a lot without knowing what Cr_0 looks like. Still, it would be good to know more. This brings us back to the questions we've discussed earlier in this chapter. Should rational priors satisfy some kind of indifference requirement? If so, what does that requirement look like? How, exactly, should priors be biased towards "uniform" worlds? Should they be aligned with some basic physical quantities?

On a more general level, we may ask how tightly priors are constrained by the norms of rationality. Some hold that there is a unique rational prior credence function. Others say that rationality is "permissive", that it allows for a wide range of priors, each of which is as rational as the other. According to the permissive view, there is an irreducibly subjective element to rational credence: perfectly rational agents with the exact same evidence may arrive at different beliefs. There may, accordingly, be no objective answer to how strongly a scientific hypothesis is supported by the evidence.

Exercise 4.9 †

Show that if a theory H entails E, and both E's prior probability is not 1, then E incrementally confirms H.

Exercise 4.10 (The raven paradox) †††

The hypothesis that all ravens are black is logically equivalent to the hypothesis that all non-black things are non-ravens. If universal generalizations are normally confirmed by their instances, and logically equivalent hypotheses are confirmed by the same data, then an observation of a white shoe ought to support the hypothesis that all ravens are black. Does it?

Essay Question 4.1

Evaluate the hypothesis that there is a unique rational prior. Assuming that beliefs evolve by conditionalising on the evidence, this is equivalent to the hypothesis that rational agents with the same evidence should have the same degrees of belief. Can you find an argument for or against this view?

Sources and Further Reading

Chapter 15 of Ian Hacking, *An Introduction to Probability and Inductive Logic* (2001) goes into some more details about conditionalization.

The cube exercise is due to Bas van Fraassen, *Laws and Symmetry* (1989, p.303). Similar problems for the Indifference Principle are often discussed under the heading of 'Bertrand's Paradox'.

The Probability Coordination Principle is best known as the 'Principal Principle', introduced in David Lewis, "A Subjectivist's Guide to Objective Chance" (1980). Lewis's formulation includes an important extra parameter for "admissible evidence" that I have omitted.

My claim that $\mathrm{Cr}(101\ G\mathrm{s}/100\ G\mathrm{s}) \approx 0.99$ is an application of Laplace's "Rule of Succession". Laplace's assumptions can be weakened. For example, we don't need to assume that you start with a uniform prior over the objective probabilities. (Search for "Bayesian convergence" if you're interested in this.)

Hempel's 1945 paper on confirmation is called "Studies in the Logic of Confirmation". It comes in two parts, and also introduces the raven paradox. Popper's falsificationist approach was first spelled out in his *The Logic of Discovery* (1935). Modern Bayesian Confirmation Theory begins with Rudolf Carnap, *Logical Foundations of Probability* (1950). Michael Strevens's Lecture Notes on Bayesian Confirmation Theory (2017) provide a good introduction. The example of the black raven in the mail is from Strevens. For a brief comparison between the frequentist ("classical") and the Bayesian approach to statistical inference, see Matthew Kotzen, "The Bayesian and Classical Approaches to statistical inference" (2022).

For an introduction to the debate over how wide the range of rational priors might be, see Christopher G. Meacham, "Impermissive Bayesianism" (2014).

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 degree of pleasure and 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. According to **utilitarianism**, an act is morally right just in case it would bring about the greatest total utility for all people. In this context, 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 different wealth levels are the same for all people (so that, for example, 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 the total utility in the population 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 a concept of utility lies in the theory of practical rationality. According to the MEU Principle, practically rational agents choose acts that maximize the credence-weighted average of the utility of the possible outcomes. 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. As a consequence, 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. The utility of a outcome measures the extent to which the agent in question wants the outcome 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 welfare, is at most a contingent fact about humans. One can easily imagine agents who are motivated by other things. We can imagine a mother who knowingly takes on hardships for the benefit of her children, or a soldier who intentionally chooses a painful death in order to save her comrades. Psychological hedonists hold 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 things other 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. On this usage, a rational agent 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 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 is raining, and you see that it is raining, 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 should not define utility as degree of pleasure or welfare. Consider a hypothetical agent who cares not just about herself, 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. But if we were to define utility as degree of pleasure or welfare, we would have to say that the agent violates the basic norm of practical rationality,

the MEU Principle.

The point generalizes. Consider a person in an abusive relationship who is manipulated into doing things that hurt or degrade 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 sanction these desires as rational or praiseworthy. 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

An outcome's utility measures the extent to which the agent is motivated to bring about the outcome. I will often say that this is the degree to which the agent *desires* the outcome, but we need to keep in mind that the word 'desire' can be misleading. For one thing, we 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, 'desire' is often associated with a particular type of motivational state. 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 comprises 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 bodily cravings, 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 need not be conscious. There is good evidence that our true motives are often not what we believe or say they are. An agent's utility function represents their 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). 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 inhibition and reinforcement, then so be it. We will model the whole network by a single utility function, staying 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 in our seemingly simple concept of utility.

Consider the following scenario.

Example 5.1 (The endowment effect)

Emily is buying a coffee mug. She is undecided between a red mug and a blue mug, and somewhat arbitrarily chooses the red one. A little later, someone offers Emily £1 if she swaps her red mug for the blue mug. Emily declines.

The kind of behaviour displayed by Emily is common. People tend to place a greater value on things they own than on things they don't own. Initially, Emily considered the two mugs equally desirable. Having bought the red mug, Emily suddenly considers it better than the blue mug.

Psychologists have offered different explanations for this effect. Some say that forgoing an owned item feels like a loss, and we don't like this feeling. Others have argued that we treat goods that we own as part of our identity; forgoing the good is thus perceived as a threat to our identity. We don't need to adjudicate between these (and other) proposals. What's important for us is that whichever explanation is correct, it should be reflected in Emily's utility function. If Emily subconsciously regards her belongings as part of her identity, and she is subconsciously motivated to preserve her identity, then her utility for an outcome that involves giving up a

previously owned good is comparatively low.

Outside philosophy – especially in economics – utility is often assumed to be a function of material goods ("commodity bundles"). On this usage, one can speak of the utility (for Emily) of *the red mug*, but one can't distinguish between, for example, the utility of *not getting the red mug* and *giving away the red mug*. No matter what utility we then assign to the two mugs, Emily's behaviour is found to violate the MEU Principle. If the red cup has greater utility than the blue cup, then Emily shouldn't have been indifferent when she decided which cup to buy. If the two cups have equal utility for Emily, then Emily should be happy to swap the red cup for the blue cup.

On our usage of 'utility', Emily's behaviour is perfectly compatible with the MEU Principle. Emily doesn't just care about which material goods she owns. She also cares about *changes* to her possessions. If she is a real person, she will also care about other things that have little to do with material goods. If we want a general model of how beliefs and desires relate to choices, we need to make room for all the desires an agent might have. We could follow the economics tradition and restrict an agent's utility function to material goods. But then we would have to add other elements to our model to account for desires that don't pertain to the possession of material goods. We will choose the theoretically simpler option of widening the definition of 'utility', so that an agent's utility function reflects everything the agent cares about. We are going to return to this theme in chapter 8.

Exercise 5.2 ††

Amartya is offered a choice between a small slice of cake, a medium-sized slice, and a large slice. He chooses the medium-sized slice. If he had been offered a choice between only the small slice and the medium-sized slice, he would have chosen the small slice.

(a) Explain why this behaviour is incompatible with the MEU Principle if the utility function is a function of material goods.

(b) Explain why the behaviour is compatible with the MEU Principle on our use of 'utility'.

Officially, we will use 'utility' to measure anything that motivates the relevant agent. It is worth pointing out, however, that our model can be usefully applied with other conceptions of utility. We might want to know, for example, what an agent

should do, *from a moral perspective*, in a situation like the Miners Problem from chapter 1, where crucial information about the world is missing. A tempting idea is that the agent should maximize expected *moral utility*, where the moral utility of an outcome is defined by some ethical theory (utilitarianism, perhaps). Similarly, a corporation's board of directors may want to know how to promote shareholder value in the light of such-and-such common information. Here the relevant utility function might be derived from the stipulated goal of promoting shareholder value, and the "credence" function might be derived from the shared information. Neither of these needs to match the beliefs and desires of any individual member of the board.

Exercise 5.3 †††

Some choices predictably change our desires. One might argue that in such a case, a rational agent should be guided not by her present desires, but by the desires she will have as a result of her choice.
Suppose you can decide right now how many drinks you will have tonight: zero, one, or two. (You have to order the drinks in advance and can't change the order at the time.) If you're sober, you prefer to have one drink rather than zero or two. But if you have a drink, you often 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. As we saw in the previous section, we need a more general conception to capture an agent's desire *not to lose a previously owned good*. We might also desire that our 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 thing, but a possible state of the world. 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. As in the case of belief, we don't distinguish between logically equivalent states of the world. If you assign high utility to drinking tea then you also assign high utility to *drinking tea or coffee but not coffee*.

Let's study how an agent's desires towards logically related propositions are related to one another. 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. You can't 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 states of increasing utility in which the difference in utility between successive states 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 both sides. 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 odd. But the

issue isn't clear. For our purposes, it proves more convenient to go with the second option. We will say that even logically necessary propositions have well-defined utility, and that their utility measures the point between "positive" and "negative" desire. If you positively want something to be the case, the utility you assign to it is greater than the utility of a tautology. If you want something not to be the case, its utility is lower than that of a tautology. Some authors make this more concrete by adopting a convention that logically necessary propositions always have utility 0.

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's clearly wrong. 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, $\mathrm{U}(A) > \mathrm{U}(B)$? Then the utility of $A \vee B$ should plausibly lie in between the utilities of A and B:

$$\mathrm{U}(A) \geq \mathrm{U}(A \vee B) \geq \mathrm{U}(B).$$

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 $\mathrm{U}(A \vee B)$ lie in between $\mathrm{U}(A)$ and $\mathrm{U}(B)$? 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 $\mathrm{U}(A \vee B)$ should plausibly be close to $\mathrm{U}(A)$. If you're confident that A is false, then $\mathrm{U}(A \vee B)$ should be near $\mathrm{U}(B)$.

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.

Here is the general rule for $\mathrm{U}(A \vee B)$, assuming A and B are incompatible. The rule was discovered by Richard Jeffrey in the 1960s and is our only basic rule of utility, apart from the assumption that logically equivalent propositions have the same utility.

Jeffrey's Axiom

If A and B are logically incompatible and $\mathrm{Cr}(A \vee B) > 0$ then

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

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 $\mathrm{Cr}(\textit{Alder Park})$ and $\mathrm{Cr}(\textit{Buckeye Park})$ 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 $\mathrm{U}(A \vee B)$ as the probability-weighted average of $\mathrm{U}(A)$ and $\mathrm{U}(B)$ *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 $(\textit{Win} \wedge \textit{Travel}) \vee (\textit{Win} \wedge \neg \textit{Travel})$, and thus has the same utility. Suppose $\mathrm{U}(\textit{Win} \wedge \textit{Travel}) = 10$, $\mathrm{U}(\textit{Win} \wedge \neg \textit{Travel}) = 0$, 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 $\mathrm{U}(\textit{Win})$?

Exercise 5.5 ††

At the beginning of this section, I argued that if $\mathrm{U}(\neg A)$ is high, then $\mathrm{U}(A)$ 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 $\mathrm{U}(A \vee \neg A) = 0$. From this assumption, and Jeffrey's axiom, it follows that $\mathrm{U}(\neg A) > 0$ just in case $\mathrm{U}(A) < 0$. More precisely, it follows that

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

Can you show how this follows?

The following consequence of Jeffrey's axiom is often useful. Assume that $S_1, \ldots, S_n$ are propositions for which it is guaranteed that exactly one of them is true. That is, any two propositions in $S_1, \ldots, S_n$ are logically incompatible (no two of the propositions can be true together), and the disjunction $S_1 \vee \ldots \vee S_n$ is logically necessary (one of the propositions must be true). Such a collection of propositions is called a **partition**. Intuitively, a partition divides the space of possible worlds into disjoint regions.

Now, Jeffrey's axiom entails that if $S_1, \ldots, S_n$ is a partition, then for any proposition A with $\mathrm{Cr}(A) > 0$,

$$\mathrm{U}(A) = \mathrm{U}(A \wedge S_1) \cdot \mathrm{Cr}(S_1/A) + \ldots + \mathrm{U}(A \wedge S_n) \cdot \mathrm{Cr}(S_n/A).$$

Let's call this the **partition formulation** of Jeffrey's axiom.

Think of A as a region in logical space. Each $A \wedge S_i$ is a disjoint subregion of A. The partition formulation says that the desirability of the whole region A is a weighted average of the desirability of the subregions, weighted by their probability conditional on A.

Exercise 5.6 †††

Derive the partition formulation of Jeffrey's axiom from Jeffrey's (original) axiom.

Exercise 5.7 ††

Derive Jeffrey's axiom from the partition formulation.

Exercise 5.8 ††

Give counterexamples to the following generalisations, assuming that an agent *desires* a proposition A iff $\mathrm{U}(A) > \mathrm{U}(\neg A)$. (Equivalently, iff $\mathrm{U}(A) > \mathrm{U}(A \vee \neg A)$.)

(a) Whenever an agent desires $A \wedge B$, they also desire AB.
(b) Whenever an agent desires A and desires B, they also desire $A \wedge B$.
(c) Whenever an agent desires A, they desire $A \vee B$.

5.4 Basic desire

I have presented Jeffrey's axiom as the sole formal requirement on rational utility. Even this much is controversial. Many 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 rational agents desire nothing but their own pleasure.) The 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 held that our basic desires 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 desires 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 disinterested in A. Here is how.

By Jeffrey's axiom,

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

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

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

Now we solve for $\mathrm{Cr}(A)$:

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

The ratio on the right-hand side is defined whenever $\mathrm{U}(A) \neq \mathrm{U}(\neg A)$, which I meant when I said that the agent is "not disinterested" in 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. 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?

Let's pretend that you have only one basic desire: to be free from pain. Let's also

pretend that this is an all-or-nothing matter. By your lights, all possible worlds in which you are free from pain are then equally good, equally desirable. In each of them, you have everything you want. The worlds in which you are *not* free from pain are also equally good. In each of them, you do not have what you want.

Let's say that a proposition has *uniform utility* for an agent if the agent does not care how the proposition is realized: all subsets of the proposition (understood as a set of possible worlds) have equal utility. In our example, being pain-free and being in pain have uniform utility.

Let's change the scenario so that you have two basic desires: being free from pain and being admired by other people. These are logically independent, so there are four combinations: (1) being pain-free and admired, (2) being pain-free and not admired, (3) being in pain and admired, and (4) being in pain and not admired. Note that these form a partition.

Being in pain no longer has uniform utility. The worlds where you are in pain divide into (better) worlds where you are in pain and admired and (worse) worlds where you are in pain and not admired. As a consequence, the utility of being in pain now depends on your beliefs: the stronger you believe that you are admired if you are in pain, the more you desire being in pain.

The four combinations of being pain-free and admired, however, have uniform utility. All worlds in which you are, say, in pain and admired are equally desirable (still pretending these are all-or-nothing matters). I will say that these combinations are your **concerns**. Intuitively, a concern is a proposition that settles everything the agent ultimately cares about. An agent's concerns always form a partition.

Remember that an outcome in a decision matrix must settle everything the agent cares about. Every outcome in every decision problem is therefore a concern. Many decision theorists use the word 'outcome' where I use 'concern'. I prefer a different label, if only because some of an agent's concern may never figure as outcomes in a decision situation.

It will be useful to have a name for an agent's utility function restricted to their concerns. I'll call it the agent's **intrinsic utility function**. (Some people say 'value function'; many just say 'utility function' and never consider the wider function we call the agent's 'utility function'.)

Formally, an intrinsic utility function assigns numbers to some partition of propositions. Intuitively, each of these propositions settles everything the agent cares about, and the numbers tell us how strongly the agent desires any particular way

of settling the things they care about. In the above example, your intrinsic utility function might by fully given as follows:

$$\begin{aligned} \mathrm{U}(\mathit{Pain} \wedge \mathit{Admired}) &= 1, \\ \mathrm{U}(\neg \mathit{Pain} \wedge \mathit{Admired}) &= 5, \\ \mathrm{U}(\mathit{Pain} \wedge \neg \mathit{Admired}) &= -5, \\ \mathrm{U}(\neg \mathit{Pain} \wedge \neg \mathit{Admired}) &= 0. \end{aligned}$$

An agent's intrinsic utility function represents the belief-independent aspect of their goals or desires. Every possible credence function is compatible with every possible intrinsic utility function.

Since no concern is ever a disjunction of other concerns, Jeffrey's axiom imposes no constraint on intrinsic utility functions. It only enters the picture when we look at the utility of propositions that aren't concerns. In effect, the axiom tells us how to derive an agent's utility for non-concerns from the agent's intrinsic utility function and their credence function. (The axiom's partition formulation makes the derivation transparent.)

In chapter 7, we will look at how an agent's intrinsic utility function might be determined by less specific desires – by a desire to be free from pain, for example, and a desire to be admired. Before we do this, we need to say more about what the utility numbers are supposed to represent. What, exactly, does it mean that a proposition has utility 5, as opposed to -5 or 27?

Exercise 5.9 †

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.

Exercise 5.10 †

Suppose your only basic desire is to that a certain person in India is happy. What does your intrinsic utility function look like?

Exercise 5.11 †††

Assume an agent's intrinsic utility function remains the same while they conditionalize on some proposition E.

(a) Can you define the new (total) utility function $\mathrm{U}_{\mathrm{new}}$ in terms of the old utility function? (That is, can you complete the equation $\mathrm{U}_{\mathrm{new}}(A) = \ldots$ in such a way that the dots make no reference to the agent's credences?)

(b) How does conditionalizing on an undesirable proposition (with $\mathrm{U}_{\mathrm{old}}(E) < \mathrm{U}_{\mathrm{old}}(\neg E)$) affect the utility of a logically necessary proposition $A \vee \neg A$?

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.

Sources and Further Reading

Chapter 6 ("Game Theory and Rational Choice") of Simon Blackburn, *Ruling Passions* (1998) eloquently defends the idea that one shouldn't constrain what rational agents may care about in the theory of practical rationality. John Broome, " 'Utility' " (1991) provides some more background and details on the two conceptions of utility.

On possible explanations for the endowment effect, see Carey K. Morewedge and Colleen E. Giblin, "Explanations of the endowment effect: an integrative review" (2015). The cake slice example is from Amartya Sen, "Internal Consistency of Choice" (1993, p.501).

The formal theory of utility in section 5.3 comes from chapter 5 of Richard Jeffrey, *The Logic of Decision* (1965/1983). The assumption that the objects of utility are the same kinds of things (propositions) as the objects of credence is common in philosophy, but not in other disciplines.

My distinction between intrinsic and non-intrinsic utility resembles a common distinction in economics between "direct utility" and "indirect utility". It also resembles the popular distinction between "intrinsic" and "instrumental" desire. But note that if A and B are concerns, then a desire for their disjunction $A \vee B$ is derived, although a disjunction is not intuitively instrumental to its disjuncts.

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 argued 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 then we can find out more about the agent's utility function than how it orders the relevant outcomes – enough to vindicate the MEU Principle. Ramsey's idea was rediscovered by John von Neumann, who published 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 foundations of modern expected utility theory.

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

Ordinalism was inspired by a wider "positivist" movement in science and philosophy. The aim of the positivists was to cleanse scientific reasoning of obscure and untestable doctrines. Every meaningful statement was to have clear conditions of verification or falsification. A hypothesis whose truth or falsity is impossible to establish by either proof or observation was to be rejected as meaningless. In psychology, this movement gave rise to **behaviourism**, the view that statements about emotions, desires, and other psychological states should be defined in terms of observable behaviour.

Today, behaviourism, and positivism more generally, have been almost entirely abandoned. In part, this is because people came to appreciate the holistic nature of scientific confirmation. Statements in successful scientific theories often have observable consequences only in conjunction with other theoretical assumptions. More practically, the behaviourist paradigm was simply 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 outdated and misguided. But 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 the 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 the state with 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. The utility of a proposition for an agent is supposed to represents the extent to which the agent, on balance, wants the proposition to be true. But what non-numerical fact about an agent makes it correct to say that their utility for a certain proposition is 5? This question still needs an answer. And there is something to be said for the idea that the answer should involve the agent's choices.

The main reason to think that an agent has such-and-such goals or desires is that

this would explain their behaviour. The point is even more obvious for the relative strength of goals or desires. I got out of bed because my sense of duty was stronger than my desire to stay in bed. Absent further explanation, the claim that my desire to stay in bed was stronger, even though I got up, is unintelligible. If we seek a standard to measure the comparative strength of different motives or goals, a natural idea is thus to look at what the agent is prepared to do.

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.

In the case of length, the conventionality of measurement essentially boils down to the choice of a **unit**. You can introduce a new measure of length simply by picking out a particular object (say, your left foot) and declare that its length is 1, with the understanding that if an object is n times as long as the chosen object then its length in your new system is n. (You could fix the unit by assigning any number greater than zero to your left foot; 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, feet, light years, and any other sensible system of representation. That my arm is four times as long as my pencil is an objective, representation-independent fact.

Temperature is different. (Or has appeared to be different until the 19th century.) People have long known that metals like mercury expand as the temperature goes up. This can be used to define a numerical representation. Imagine we put a certain amount of mercury in a narrow glass tube. The higher the temperature, the more

of the glass tube is filled with the expanding mercury. To get a numerical measure of temperature, we now need to mark *two* points on the tube, a unit and a **zero**. We could, for example, mark the point at which water freezes as 0 and the point at which it boils as 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.

The Celsius scale for temperature and the Fahrenheit scale have different units and zeroes. As a result, 10 degrees Celsius is 50 degrees Fahrenheit, and 20 degrees Celsius is 68 degrees Fahrenheit. The ratio between the two temperatures is not preserved, so these scales are not ratio scales. Scales in which both the zero and the unit are a matter of convention are called **interval scales**.

Exercise 6.1 ††

Someone might suggest that we only need to mark a unit on the glass tube, since we are effectively measuring the volume of the mercury in the tube, and volume has a ratio scale: zero simply means that the mercury fills up none of the tube. Does this show that temperature has a (natural) ratio scale?

Ratio scales and interval scales are both called **cardinal** scales, in contrast to **ordinal scales**, in which the only thing that is not conventional is which of two objects is assigned a greater number.

The ordinalists held that utility has only an ordinal scale (hence the name of the movement). All we have to go by in order to measure utilities, the ordinalists assumed, are the agent's choices. If you choose tea over coffee and coffee over milk, we may infer that your utility for tea is greater than your utility for coffee, which in turn is greater than your utility for milk. But any assignment of numbers that respects this ordering is as correct as any other. We could say that for you, tea has utility 3, coffee 2, and milk 1, but we could equally say that tea has utility 100, coffee 0, and milk -8.

If the ordinalists were right, then whether an act in a decision problem maximizes expected utility would often depend on arbitrary choices in the measurement of utility. The MEU Principle would be indefensible. If, on the other hand, utility has an interval scale, then different measures of utility never disagree on the ranking of acts in a decision problem. A ratio scale is not required.

Exercise 6.2 †

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 6.3 ††

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, $\mathrm{U}(A) = x \cdot \mathrm{U}'(A) + y$. 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 $\neg S$.)

If we want to rescue the MEU Principle from ordinalist skepticism, we therefore don't need to explain what makes it the case that your utility for tea is 3 rather than 100. We can accept that the exact numbers are a conventional matter of representation. Nor do we need to explain what makes your utility for tea twice your utility for coffee; such ratios also need not track anything real. But we do have to explain why, if we arbitrarily mark your utility for tea as (say) 1 and your utility for coffee as 0, then your utility for milk is fixed at a particular value: why it has to be -1 (say), rather than -7, even though both hypotheses appear to be in line with your choices.

6.3 Utility from preference

I am now going to describe John von Neumann's method for determining an agent's utility function from their preferences or choice dispositions. More precisely, what we are going to determine is the agent's *intrinsic* utility function. Recall from section 5.4 that an intrinsic utility function assigns a utility to each of the agent's *concerns*,

where a "concern" is a proposition that settles everything the agent ultimately cares about.

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 ultimately 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 determine the utility the agent assigns to $a, b, c, \ldots$.

We do this by looking at the agent's **preferences**, which we assume to represent their choice dispositions. For example, if the agent would choose reward a when given a choice between a and b, we say that the agent prefers a to b. The ordinalists did not challenge the assumption that people have preferences.

Let's introduce some shorthand notation:

$$A \succ B \Leftrightarrow \text{The agent prefers } A \text{ to } B.$$
$$A \sim B \Leftrightarrow \text{The agent is indifferent between } A \text{ and } B.$$
$$A \succsim B \Leftrightarrow A \succ B \text{ or } A \succsim B.$$

(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 will use facts about the agent's preferences to construct an intrinsic utility function U (an assignment of numbers to rewards) that **represents** the agent's preferences, in the sense that for all rewards a and b, $\mathrm{U}(a) > \mathrm{U}(b)$ iff $a \succ b$, and $\mathrm{U}(a) = \mathrm{U}(b)$ iff $a \sim b$.

Let's begin. We accept that the choice of unit and zero is a matter of convention, so we take arbitrary rewards a and b such that $b \succ a$ and set $\mathrm{U}(a) = 0$ and $\mathrm{U}(b) = 1$. This resembles the conventional choice of using 0 for the temperature at which water freezes and 100 for the temperature at which it boils.

Exercise 6.4 ††

If our agent is indifferent between all rewards, then the 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, we can now 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 L is

$$\mathrm{EU}(L) = 1/2 \cdot \mathrm{U}(a) + 1/2 \cdot \mathrm{U}(b) = 1/2 \cdot 0 + 1/2 \cdot 1 = 1/2.$$

If the agent obeys the Probability Coordination Principle and the MEU Principle, and she is indifferent between L and c, we can infer that c has utility $1/2$.

Exercise 6.5 †

Suppose $\mathrm{U}(a) = 0, \mathrm{U}(b) = 1$, and $\mathrm{U}(c) = 1/2$. Draw a decision matrix representing a choice between c and L, and verify that the two options have equal expected utility.

Exercise 6.6 ††

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

If the agent isn't indifferent between c and L, we try other lotteries, until we find one the agent regards as equally good as c. For example, suppose the agent is indifferent between c and a lottery L' that gives them a with probability $4/5$ and b with probability $1/5$. Since the expected utility of this lottery is $1/5$, we could 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? In this case, we 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 lottery L'' that gives them either a or c with equal probability, then

c must have utility 2. That's because the expected utility of L'' is

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

Since the agent is indifferent between L'' and b, which has a guaranteed utility of 1, the lottery must have expected utility 1. So $1 = 1/2 \cdot U(c)$. And so $\mathrm{U}(c) = 2$. 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 $\mathrm{U}(c) = 1/x$.

Exercise 6.7 ††

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 their preferences between rewards and lotteries. The resulting (intrinsic) utility function has an arbitrary unit and zero, but once these are fixed, the other utilities are no longer an arbitrary matter of convention. We have a cardinal utility scale. We have answered the ordinalist challenge. Or so it seems.

6.4 The von Neumann and Morgenstern axioms

The method described in the previous section assumes that the agent obeys the MEU Principle. This may seem strange. The ordinalists argued that the MEU Principle makes 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 agent's beliefs and desires and try to find out the optimal choices. Now we start with the agent's choices and try to find out the agent's 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 Principle, the MEU Principle tells us what lotteries an agent should be disposed to accept if she has a given utility function. If she doesn't accept these lotteries, we can infer that she doesn't have the utility function.

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?

Not quite. It's true that *if the method works*, then the agent will evaluate lotteries by their expected utility, relative to the utility function identified by the method. But the method is not guaranteed to work. Nor does it settle how the agent evaluates the options in decision problems in which the relevant objective probabilities are unknown.

Here is one way in which the method might fail to work. We have assumed that if an agent ranks some reward c as between a and b, then the agent is indifferent between c and some lottery between a and b. This 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. To state these conditions, we 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

If $A \succ B$ and $B \succ C$ then $A \succ C$; if $A \sim B$ and $B \sim C$ then $A \sim C$.

Continuity

If $A \succ B$ and $B \succ C$ then there are lotteries L_1 and L_2 between A and C such that $A \succ L_1 \succ B$ and $B \succ L_2 \succ C$.

Independence (of Irrelevant Alternatives)

If $A \succsim B$, 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 $L_1 \succsim L_2$.

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, that represents the agent's preferences (in the sense that $\mathrm{U}(A) > \mathrm{U}(B)$ iff $A \succ B$, and $\mathrm{U}(A) = \mathrm{U}(B)$ iff $A \sim B$). Von Neumann and Morgenstern also proved that the function U is unique except for the choice of unit and zero: any two functions U and U' that represent the agent's preferences differ at most in the choice of unit and zero. These two results are known as the **von Neumann-Morgenstern Representation Theorem**.

If we adopt von Neumann's method for measuring an agent's utilities in terms of their choice dispositions, then the MEU Principle for choices involving lotteries is automatically 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 reduces to the claim that their preferences ought to satisfy the conditions. For this reason, the conditions are often called the **axioms of expected utility theory**.

Von Neumann therefore discovered not only a response to the ordinalist challenge (at least for agents who satisfy the axioms). He also discovered 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 lotteries by their expected utility.
3. This function U is the agent's true utility function.
4. Therefore: A rational agent ranks lotteries by their expected utility.

Exercise 6.8 †

Maurice would 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. If he were offered a choice between going to the mountains 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 asked how an agent's credences could be measured or defined. The betting interpretation gave a simple answer, but we found that it relies on implausible assumptions about the agent's utility function. In the meantime, we have learned from von Neumann how we might derive an agent's intrinsic utilities from their choice dispositions. With this information in hand, we might try again to determine the agent's credence function by offering them suitable bets.

Frank Ramsey, way ahead of his time in 1926, showed how the two tasks can be combined. He described a method for determining both a credence function and a utility function from an agent's preferences.

Ramsey begins by determining the agent's utility function. We already know one way of doing this – von Neumann's. Ramsey's method is a little different, and worth going over.

Ramsey doesn't use lotteries. Instead, he uses deals whose outcome depends on some proposition the agent doesn't intrinsically care about. Suppose N is a proposi-

tion whose truth-value you don't care about (say, that the number of stars is even). Suppose also that your credence in N is $1/2$. Instead of offering you a lottery that yields outcome a or outcome b with equal chance, we can then offer you a deal that leads to a if N and to b if $\neg N$.

I will refer to conditional deals of the form 'A if X, B if $\neg X$' as **gambles**. Notice that every act in every decision problem corresponds to a (possibly nested) gamble. In the mushroom problem from chapter 1, for example, eating the mushroom amounts to choosing the gamble '*Dead* if *Poisonous*, *Satisfied* if not *Poisonous*'; not eating the mushroom amounts to choosing '*Hungry* if *Poisonous*, *Hungry* if not *Poisonous*'.

We need to identify a suitable proposition N with credence $1/2$, merely by looking at an agent's preferences.

Let's say that a proposition A is *neutral* for an agent if, for any conjunction of rewards R, the agent is indifferent between $R \wedge A$ and $R \wedge \neg A$. Intuitively, a neutral proposition is one the agent doesn't care about. Now let a and b be two rewards such that $a \succ b$. Suppose we find a neutral proposition N such that the agent is indifferent between the gambles 'a if N, b if $\neg N$' and 'b if N, a if $\neg N$'. If the agent ranks gambles by their expected utility, we can infer that the two gambles have equal expected utility:

$$\mathrm{Cr}(N) \cdot \mathrm{U}(a) + \mathrm{Cr}(\neg N) \cdot \mathrm{U}(b) = \mathrm{Cr}(N) \cdot \mathrm{U}(b) + \mathrm{Cr}(\neg N) \cdot \mathrm{U}(a).$$

Assuming that the agent's credences are probabilistic, so that $\mathrm{Cr}(\neg N) = 1 - \mathrm{Cr}(N)$, it follows that $\mathrm{Cr}(N) = 1/2$. (As you may check.)

We now use this proposition N to determine the agent's utility function.

As before, we fix the unit and zero by taking arbitrary rewards with $b \succ a$ and set $\mathrm{U}(a) = 0$ and $\mathrm{U}(b) = 1$. Then we go through all the rewards until we find one for which the agent is indifferent between c and the gamble 'a if N, b if $\neg N$'. Since this gamble has expected utility $1/2$, we can infer that c has utility $1/2$.

In the next step, we can use gambles involving a, b, and c to determine the utility of further rewards. For example, if the agent is indifferent between a reward d and the gamble 'a if N, c if $\neg N$', then d must have utility $1/4$. And so on.

We can also determine the utility of rewards that don't lie between a and b. Suppose, for example, that a reward e is preferred to b, and the agent is indifferent between the gambles 'a if N, e if $\neg N$' and 'c if N, b if $\neg N$', where c is the earlier

reward whose utility we've determined to be $1/2$. Then the utility of e must be 1.5.

Exercise 6.9 †

Explain this last claim. That is, show that if $\mathrm{U}(a) = 0, \mathrm{U}(b) = 1, \mathrm{U}(c) = 1/2$, and the agent evaluates gambles by their expected utility, then they are indifferent between 'a if N, e if $\neg N$' and 'c if N, b if $\neg N$' only if $\mathrm{U}(e) = 1.5$.

If all went well, we now know the utility the agent assigns to all rewards. We still need to determine the agent's credence in propositions other than N.

Let X be some proposition whose credence we want to determine. Ramsey instructs us to find rewards a, b, and c such that the agent is indifferent between a and the gamble 'if X then b, if $\neg X$ then c'. The gamble's expected utility is $\mathrm{Cr}(X) \cdot \mathrm{U}(b) + \mathrm{Cr}(\neg X) \cdot \mathrm{U}(c)$. Since the agent is indifferent between the gamble and a, we can infer that

$$\mathrm{U}(a) = \mathrm{Cr}(X) \cdot \mathrm{U}(b) + (1 - \mathrm{Cr}(X)) \cdot \mathrm{U}(c).$$

Solving for $\mathrm{Cr}(X)$ yields

$$\mathrm{Cr}(X) = \frac{\mathrm{U}(a) - \mathrm{U}(c)}{\mathrm{U}(b) - \mathrm{U}(c)}.$$

All quantities on the right-hand side are known. We have determined $\mathrm{Cr}(X)$.

Like von Neumann's method, Ramsey's method only works if the agent's preferences satisfy certain formal conditions or "axioms". Ramsey lists eight axioms, the details of which won't be important for us.

Ramsey's Representation Theorem states that if (but not only if) an agent's preferences satisfy his eight conditions, then there is a utility function U and a probability function Cr which together represent the agent's preferences, in the sense that (i) $A \succ B$ iff the expected utility of A, relative to Cr and U, is greater than that of B, and (ii) $A \sim B$ iff A and B have equal expected utility. The theorem also says that Cr is unique and U is unique expect for the choice of zero and unit.

What can this do for us? Ramsey's idea is that we may *define* an agent's credence and (intrinsic) utility as whatever functions Cr and U "make sense of their preferences". By this I mean that the agent prefers some proposition A to a proposition B iff the expected utility of A, computed with Cr and U, is greater than that of B. I also assume that in order for Cr to "makes sense" of the agent's preferences, it must

conform to the rules of probability.

Ramsey's Representation Theorem assures us that if the agent's preferences satisfy his axioms, then *there are* functions Cr and U that make sense of the agent's preferences. Moreover, while there are different such pairs of functions Cr and U, they all involve the exact same function Cr, and the different U functions differ only in their choice of unit and zero. For agents who satisfy the axioms, our definition is therefore guaranteed to identify a unique credence function and a utility function that is determinate enough to vindicate the MEU Principle.

If we could convince ourselves that Ramsey's axioms are requirements of rationality, Ramsey's approach would deliver a more comprehensive argument for the MEU Principle than what 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, people often face options for which they don't know the objective probability of the outcomes. Why should they rank such options by their expected utility? On Ramsey's approach, the only way they could fail to do so is by violating at least one of the axioms.

Ramsey's approach also suggests a new argument for probabilism, the claim that rational degrees of belief conform to the rules of probability. (This was Ramsey's actual aim.) Again, the requirement reduces to the preference axioms. On the proposed definition of credence, any agent who obeys the axioms automatically has probabilistic credences. If you don't have probabilistic credences, you violate the axioms.

Exercise 6.10 ††

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

Unfortunately, Ramsey's axioms can hardly be considered requirements of rationality. Note, for example, that his method doesn't work unless there is a neutral proposition N with credence $1/2$, or unless there is a reward c for which the agent is indifferent between c and the gamble 'a if N, b if $\neg N$'. Ramsey's axioms 1 and 6 ensure that these conditions are met, but it is hard to see why they should be requirements of rationality.

Later authors have improved upon Ramsey in some respects. They have come up

with other (and generally more complicated) methods for determining credences and utilities from preferences. The best known of these proposals is due to Leonard Savage, published in his *Foundations of Statistics* (1954) – the second-most influential book in the history of decision theory, after *Game Theory and Economic Behaviour*. I won't go through Savage's method and axioms. Suffice it to say that his axioms still include conditions that nobody can seriously regard as requirements of rationality, let alone as requirements that anyone must meet in order to have credences and utilities.

6.6 Preference from choice?

Von Neumann and Ramsey both take as their starting point an agent's preferences, represented by the relations $\succ$, $\sim$, and $\succsim$. I suggested that we might read '$A \succ B$' as saying that the agent would choose A if given a choice between A and B. On this interpretation, von Neumann and Ramsey showed how we might determine an agent's utilities (and credences, in Ramsey's case) from their choice dispositions, assuming that these dispositions satisfy certain conditions ("axioms").

Let's be clear why I talk about dispositions. An agent's *dispositions* reflect what the agent *would* do if such-and-such circumstances were to arise. There is little hope of determining an agent's utilities or credences from their actual choices alone. Von Neumann and Ramsey certainly appeal to all sorts of choices most real agents never face.

Exercise 6.11 ††

Suppose we define '$A \sim B$' as 'the agent has faced a choice between A and B and expressed indifference', and '$A \succ B$' as 'the agent has faced a choice between A and B and expressed a preference for A. Which of the von Neumann and Morgenstern axioms then become highly implausible (no matter what exactly we mean by "expressing" indifference or preference)?

Now one of the problems for the betting interpretation, from section 3.4, returns with a vengeance. If an agent is not facing a choice between two options A and B, then offering them the choice would change their beliefs. Among other things, the agent would come to believe that they face that choice. From the fact that the

agent *would* choose (say) A if they *were* offered the choice, we can't infer that the agent's *actual* expected utility of A is greater than that of B, even if we assume that the agent obeys the MEU Principle. Expected utilities depend on credences, and perhaps A only has greater expected utility after the agent's credences are updated by the information that they can choose between A and B.

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 $3/4$. Ramsey's method might similarly look at your preferences between peace in Syria and gambles like 'peace in Syria if the number of stars is even, being able to play the piano if the number is odd'. If you thought you'd face this bizarre choice, your beliefs would surely be quite different from your actual beliefs. (Indeed, merely from being offered the choice, you could figure out that either there is peace in Syria or you can play the piano.)

Even in the rare case where an agent actually faces a relevant choice between A and B, we arguably can't infer that whichever option they choose (say, A) has greater expected utility.

For one thing, the agent might be indifferent between A and B, and have chosen A at random. Choice dispositions arguably can't tell apart $A \succ B$ and $A \sim B$. The agent might also be mistaken about their options. If I offer you a choice between an apple and a banana, and you falsely believe that the banana is a wax banana, your choice of the apple doesn't show that you prefer an apple over a (real) banana. You might be similarly mistaken about which gambles or lotteries are on offer.

The upshot is that we need to distinguish (at least) two notions of preference. One represents the agent's choice dispositions: whether they would choose A over B in a hypothetical situation in which they face this choice. The other represents the agent's current ranking of rewards and gambles or lotteries: whether by the lights of the agent's current beliefs and desires, A is better than B. Von Neumann and Ramsey have at best shown how to derive utilities and credences from preferences in the second sense.

This could still be valuable. We might still get an interesting argument for probabilism and the MEU Principle. Moreover, there is plausibly *some* connection be-

tween preference in the second sense and choices dispositions. We haven't fully solved the measurement problem for credences and utilities. But one might hope that we are at least a few steps closer.

Essay Question 6.1

An agent's choice dispositions provide information about their beliefs and desires, but perhaps it is a mistake to think that one can determine the agent's beliefs and desires by looking at nothing but their choice dispositions. What other facts about the agent might one take into account? Evaluate the prospects of measuring an agent's utilities and/or credences based on these other facts, perhaps in combination with the agent's choice dispositions.

Sources and Further Reading

The 1926 draft in which Ramsey shows how one might derive utilities and credences from preferences is called "Truth and Probability". Edward Elliott, "Ramsey without Ethical Neutrality: A New Representation Theorem" (2017) provides a useful summary and suggests some improvements to Ramsey's method.

For a good discussion of Savage's approach and its limitations, see chapter 3 of James Joyce, *The Foundations of Causal Decision Theory* (1999). A useful, but mathematically heavy, survey of other representation theorems in the tradition of Ramsey, Savage, and von Neumann and Morgensterm is Peter Fishburn, "Utility and Subjective Probability" (1994).

Preference-based approaches to utility are standard in economics, but fairly unpopular in philosophy. Christopher J.G. Meacham and Jonathan Weisberg, "Representation theorems and the foundations of decision theory" (2011) lists some common philosophical misgivings.

On the connection between preference and choice behaviour, see, for example, chapter 3 of Daniel M. Hausman, *Preference, Value, Choice, and Welfare*, and Johanna Thoma, "In defence of revealed preference theory" (2021).

The Maurice exercise is from John Broome, *Weighing Goods* (1991, p.101).

7 Separability

7.1 The construction of utility

When a possible outcome looks attractive, then this is usually because it has attractive aspects. It may also have unattractive aspects, but the attractive aspects (the "pros") outweigh the unattractive aspects (the "cons"). In this chapter, we will explore how this weighing of different aspects might work.

Take a concrete example. You are looking for a flat to rent. There are two options. *A* is a small and central flat that costs £800/month. *B* is a larger flat in the suburbs for £600/month. You might draw up a lists of pros and cons for each option, and give them a weight, like so:

A	*B*
good location (+2)	bad location (-2)
a little small (-1)	good size (+3)
expensive (-3)	a little expensive (-1)

You might then determine the *total* utility of each option as the Asum of these numbers, so that U(A) is +2-1-3 = -2, while U(B) is -2+3-1 = 0.

Is this a reasonable approach? It looks OK in this example. But we have to be careful. Suppose you had drawn up the following table.

A	*B*
good location (+2)	bad location (-2)
short commute (+1)	long commute (-1)
can get up later (+1)	have to get up earlier (-1)
a little small (-1)	good size (+3)
expensive (-3)	a little expensive (-1)

Now $\mathrm{U}(A)$ comes out as 0 and $\mathrm{U}(B)$ as -2. Do you see what's wrong with this table?

The problem is that the first three criteria in the list aren't independent. Once you've taken "good location" into account, you shouldn't *additionally* take into account "short commute" and "can get up later". Location, size, and costs are independent criteria. Location and commute time are not.

But what, exactly, does independence mean here? There is no *logical* connection between "good location" and "short commute". And there may well be a strong statistical connection between (say) location and costs.

7.2 Additivity

Let's stick with the flat example. We assume that you care about certain aspects of a flat: size, location, and costs. We'll call these aspects **attributes**. Let's assume that size, location, and costs are all the attributes that ultimately matter to you. Your preferences between possible flats is then determined by your preferences between combinations of these attributes. If two flats perfectly agree in each of the three attributes then you are always indifferent between them. If you prefer one flat to another, that's always because you prefer the combined attributes of the first to those of the second.

Instead of talking about the desirability of a particular flat, we can therefore talk about the desirability of its attributes. We'll write combinations of attributes as lists enclosed in angular brackets. '$\langle 40\mathrm{m}^2, \text{central}, £500\rangle$', for example, would represent any flat with a size of 40 m^2, central location, and monthly costs of £500. We are interested in the utility you assign to any such list.

Strictly speaking, of course, utility functions don't assign numbers to lists, or even to flats. When I say that you prefer one kind of flat over another, what I really mean is that you prefer living in the first kind of flat over living in the other. In full generality, we should speak about attributes of worlds, not of flats. To keep things simple, we currently assume that the only thing you ultimately care about is what kind of flat you are living in (or going to live in). A list like $\langle 40\mathrm{m}^2, \text{central}, £500\rangle$ therefore settles everything you ultimately care about. It represents one of your "concerns", in the terminology of section 5.4.

In the example from section 5.4, we assumed that you care about two things: being free from pain and being admired. We pretended that these are all-or-nothing matters.

The resulting four concerns could be represented by the following lists:

$$\langle \mathit{Pain}, \mathit{Admired} \rangle, \langle \neg \mathit{Pain}, \mathit{Admired} \rangle, \langle \mathit{Pain}, \neg \mathit{Admired} \rangle, \langle \neg \mathit{Pain}, \neg \mathit{Admired} \rangle.$$

Here, there are two attribute, each of which can take two value. The first attribute specifies whether you are in pain, and the answer is either yes or no. The second attribute similarly specifies whether you are admired. If we allowed for different degrees of pain, then the first attribute would have more than two possible values. We could, for example, distinguish $\langle \mathit{Little\ Pain}, \mathit{Admired} \rangle$ from $\langle \mathit{Strong\ Pain}, \mathit{Admired} \rangle$.

In the flat example, we have three attributes, each of which can take many different values: size, location, and costs. Your intrinsic utility function assigns a desirability score to all possible combinations of these values.

If you're like most people, we can say more about how these scores are determined. For example, you probably prefer cheaper flats to more expensive flats, and larger flats to smaller flats. The "weighing up pros and cons" idea suggests that the overall score for a flat is determined by adding up individual scores for the flat's properties. Let's spell out how this might work.

We want to compute the utility of any given attribute list as the sum of numbers assigned to the elements in the list. We'll call these numbers **subvalues**. A size of 40 m^2 might have subvalue $V_S(40\,\mathrm{m}^2) = 1$. Central location might have subvalue $V_L(\text{central}) = 2$. Monthly costs of £500 might have subvalue $V_C(£500) = -1$. Note that we have three different subvalue functions: one for size, one for location, one for costs. The overall value (utility) of $\langle 40\,\mathrm{m}^2, \text{central}, £500 \rangle$ would then be the sum of these subvalues:

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

If U is determined by adding up subvalues in this manner, then it 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, we could introduce scaling factors w_S, w_L, w_C, so that

$$\mathrm{U}(\langle 40\,\mathrm{m}^2, \text{central}, £500 \rangle) = w_s \cdot V_S(40\,\mathrm{m}^2) + w_L \cdot V_L(\text{central}) + w_C \cdot V_C(£500).$$

For convenience, we will omit the weights by folding them into the subvalues. We

will let $V_S(200\,\mathrm{m}^2)$ measure not just how awesome it would be to have a 200 m^2 flat, but also how important this feature is compared to cost and location.

Subvalue functions are typically defined over propositions that don't have uniform utility. Recall that, strictly speaking, '200 m^2' expresses the proposition that you are going to live in a 200 m^2 flat. Some of the worlds where you live in such a flat are great. Others are bad. That's because you also care about location and costs, and the 200 m^2 worlds differ in these respects. An (improbable) world in which you rent a 200 m^2 central flat for £100/month is better than a (more probable) world in which you rent a 200 m^2 flat in the suburbs for £1000/month. As a result, the utility of 200 m^2 may be low, even though the subvalue is high.

Informally, the *utility* of 200 m^2 measures the desirability of the relevant proposition. Would you be glad to learn that you are going to rent a 200 m^2 flat? Perhaps not, because the large size indicates high costs and bad location. The *subvalue* of 200 m^2 is not sensitive to your beliefs. It measures the intrinsic desirability of that size, no matter what it implies or suggests about other attributes. It measures how much a size of 200 m^2 contributes to the overall desirability of a flat, holding fixed the other attributes.

Exercise 7.1 †††

We could define a concept of additivity purely in terms of utility. Let's say that a utility function U is *utility-additive* with respect to attributes $A_1, \ldots, A_n$ iff $\mathrm{U}(\langle A_1, \ldots, A_n \rangle) = \mathrm{U}(A_1) + \ldots + \mathrm{U}(A_n)$. Explain why your utility function in the flat example isn't utility-additive with respect to size, location, and costs.

Exercise 7.2 ††

Additivity greatly simplifies an agent's psychology. Suppose an agent's basic desires pertain to 10 logically independent 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 intrinsic utility function in a database, we would therefore need to store up to 1024 numbers. How many numbers do we need to store in the database if the agent's intrinsic utility function is additive?

7.3 Separability

Under what conditions is intrinsic utility determined by adding subvalues? How are different subvalue functions related to one another? We can get some insight into these questions by following an idea from the previous chapter and study how intrinsic utility might be derived from preferences.

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 rather than 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 the overall ranking of attribute lists.

So assume we have preference relations $\succ$, $\succsim$, $\sim$ between lists of attributes. (We aren't interested in lotteries or gambles this time, only in complete concerns.) 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 we have

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

If, like most people, you prefer to pay less rather than more, then your subvalue function V_C is a decreasing function of monthly costs: the higher the costs c, the lower $V_C(c)$. This doesn't mean that you 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. But the following should hold: whenever two flats agree in size and location, and one is cheaper than the other, then you prefer the cheaper one.

Let's generalize this idea.

Consider an attribute list $\langle A_1, A_2, \ldots A_n\rangle$, and let A_1' be an alternative to A_1. If, for example, the first position in an attribute list represents monthly costs, then A_1 might be £400 and A_1' £500. We can now compare $\langle A_1, A_2, \ldots A_n\rangle$ to $\langle A_1', A_2, \ldots A_n\rangle$ – a hypothetical flat that's like the first in terms of size and location, but costs £100 more. If

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

we say that you prefer A_1 to A_1' *conditional on* $A_2, \ldots, A_n$.

Suppose you prefer A_1 to A_1' conditional on any way of filling in the remainder

$A_2, \ldots, A_n$ of the attribute list. In that case, we can say that your preference of A_1 over A_1' is *independent* of the other attributes.

In the flat example, your preference of £400 over £500 is plausibly independent of the other attributes: whenever two possible flats agree in size and location, but one costs £400 and the other £500, you plausibly prefer the one for £400. (We are still assuming that size, location, and costs are all you care about.)

We can similarly consider alternatives A_2 and A_2' that may figure in the second position of an attribute list, and alternatives A_3 and A_3' in the third positions, and so on. If we find that your preferences between A_i and A_i' are always independent of the other attributes, we say that your preferences between attribute lists are **weakly separable**.

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. But we don't have a violation of weak separability, simply because the relevant attribute lists differ in more than one position.

That's why weak separability is called 'weak'. To rule out the present kind of dependence, we need to strengthen the concept of separability. Preferences are called **strongly separable** if the ranking of lists that differ in *one or more positions* does not depend on the attributes in the remaining positions, in which they do not differ. 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. Your preferences aren't strongly separable.

(Are they weakly separable? We can't say. I have only specified how you rank two pairs of lists. Your preferences are presumably defined for many other combinations of flat size, location, and costs. There's no violation of weak separability in the two data points I have given. But there might be a violation elsewhere.)

Exercise 7.3 ††

Suppose all you care about is the degree of pleasure of you and your three friends, which we can represent by a list like $\langle 10, 1, 2, 3 \rangle$. Suppose further that you prefer states in which you all experience equal pleasure to states in which your degrees of pleasure are 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, based on the information provided? Which violate strong separability?

(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 are a function U, assigning numbers to the lists, and subvalue functions $V_1, V_2, \ldots, V_n$, assigning numbers to the items on the lists, such that the following two conditions are satisfied. First, the preferences are represented by U. That is, for any two lists A and B,

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

Second, the U-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:

$$\mathrm{U}(\langle A_1, A_2, \ldots, A_n \rangle) = V_1(A_1) + V_2(A_2) + \ldots + V_n(A_n).$$

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 stronger condition called 'double-cancellation' is required. 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.

Debreu's theorem has an interesting corollary. Suppose a utility function U has an additive representation in terms of certain attributes. One can show that if the attributes are sufficiently fine-grained, and small differences to the attributes make for small difference in overall utility, then every utility function U' that has an additive representation in terms of the relevant attributes differs from U at most in the choice of unit and zero.

This suggests a new 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 rational (intrinsic) preferences should be strongly separable and that an adequate representation of such preferences should involve an additive utility function. The only arbitrary aspect of a utility representation would then be the choice of unit and zero.

Exercise 7.5 ††

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

$$\mathrm{U}(\langle A_1, A_2, \ldots, A_n \rangle) = V_1(A_1) + V_2(A_2) + \ldots + V_n(A_n).$$

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

functions $V_1', V_2', \ldots, V_n'$ such that

$$\mathrm{U}'(\langle A_1, A_2, \ldots, A_n \rangle) = V_1'(A_1) + V_2'(A_2) + \ldots + V_n'(A_n).$$

Exercise 7.6 †††

Assume all you care about are your wealth and your height. On one way of representing your preferences, the utility you assign to any combination of wealth w (in GBP) and height h (in meters) is $\mathrm{U}(\langle w, h \rangle) = w \cdot h$. Do your preferences have an additive representation? Explain your answer.

Why might one think that rational preferences should be separable? Remember that we are talking about preferences over "attribute lists" that settle everything the agent ultimately cares about, with each position in a list settling one question that intrinsically matters to the agent. In our toy example, these were the size, location, and costs of their flat. More realistically, items in the attribute list might be the agent's level of happiness, their social standing, the well-being of their relatives, etc. Now, if an agent has a basic desire for, say, happiness, then we would expect that increasing the level of happiness, while holding fixed everything else the agent cares about, always is a change for the better. That is, if two worlds w_1 and w_2 agree in all respects that matter to the agent except that the agent is happier in w_1 than in w_2, then we would expect the agent to prefer w_1 over w_2. From this perspective, separability might be understood as a condition on how to identify basic desires: if an agent's preferences over some attribute lists are not separable, then the attributes don't represent (all) the agent's basic (intrinsic) desires.

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. The hedonist conjecture leaves open how people rank different ways pleasure can be distributed over a lifetime. Unless an agent just cares about their 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 intrinsic utility function. The hedonist's intrinsic utility 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 utility 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 utility 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 $\langle p_1, p_2, \dots, p_n \rangle$ than in $\langle p'_1, p'_2, \dots, p'_n \rangle$, then there is also more pleasure in $\langle p_0, p_1, p_2, \dots, p_n \rangle$ than in $\langle p_0, p'_1, p'_2, \dots, p'_n \rangle$.

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 function of the form

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

where δ is a fixed number greater than 1. The interesting thing here is that the subvalue function for any time equals the subvalue function V_1 for the first time, scaled by an exponential **discounting factor** δ^i.

If a hedonist has strongly separable and stationary preferences, then her preferences over time streams are fixed by two things: how much she values present pleasure, and how much she discounts the future. If $\delta = 1$, the agent values pleasure equally, no matter when it occurs. If $\delta = 1/2$, then one unit of pleasure tomorrow is worth half as much as to the agent as one unit today; the day after tomorrow it is worth a quarter; 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 the relevant events (talking to friends, etc.) right now and by 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.

An argument in favour of stationarity is that it is often thought to be required to pro-

tect the agent from a kind of disagreement with her future self. To illustrate, suppose you prefer $\langle 10, 0, 0, 0, \ldots \rangle$ to $\langle 0, 11, 0, 0, \ldots \rangle$ because you care more about today's pleasure than about tomorrow's. You care less about the difference between getting pleasure in four days and getting it in five days, so you prefer $\langle 0, 0, 0, 0, 11, 0, 0, \ldots \rangle$ to $\langle 0, 0, 0, 11, 10, 0, 0, \ldots \rangle$. These preferences violate stationarity. Stationarity would imply that if you prefer $\langle 10, 0, 0, 0, \ldots \rangle$ to $\langle 0, 11, 0, 0, \ldots \rangle$ then you also prefer the first stream to the second if both are prefixed with 0, and therefore also if both are prefixed with two 0s, and with three 0s. Now suppose your (non-stationary) preferences remain the same for the next 4 days. At the end of this time, you'd still rather have 10 units of pleasure today than 11 tomorrow: you still prefer $\langle 10, 0, 0, 0, \ldots \rangle$ to $\langle 0, 11, 0, 0, \ldots \rangle$. But your "today" is what used to be "in 4 days". Your new preferences disagree with those of your earlier self, in the sense that the worlds your former self regarded as better you now regard as worse. This kind of disagreement is called **time inconsistency**.

Empirical studies suggest that time inconsistency is pervasive. People often prefer their future selves to study, eat well, and exercise, but choose burgers and TV for today.

These preferences do look problematic. Other apparent violations of stationarity, and even separability across time, however, look OK. Suppose you like to have a glass of wine every now and then. But only now and then; you don't want to have wine every day. It seems to follow that your preferences violate both separability and stationarity. You violate stationarity because even though you might prefer a stream $\langle$wine, no wine, no wine, $\ldots \rangle$ to $\langle$no wine, no wine, no wine, $\ldots \rangle$, 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 of pleasure over wildly fluctuating levels, or the other way round.

One might argue, however, that in these cases the items in the time streams do not represent you basic desires, or not all of them. If, for example, you have a preference for constant levels of pleasure, then your basic desires don't just pertain to how much pleasure you have today, how much pleasure you have tomorrow, and so on. You have a further basic desire: that your pleasure be constant from day to day.

Exercise 7.8 ††

Are your preferences in the wine example time-inconsistent, in the sense that what you prefer for your future self is not what your future self prefers for itself?

Exercise 7.9 ††

If you care about whether you have wine on consecutive days, then arguably an adequate time stream for your concerns shouldn't simply specify, for each day, whether you do or do not have wine, but also whether you are *having wine after having had wine the previous day*. An adequate representation of a week in which you have wine on days 2, 4, and 5 would therefore be $\langle \bar{W}\bar{P}, W\bar{P}, P\bar{W}, W\bar{P}, WP, \bar{W}P, \bar{W}\bar{P} \rangle$, where W means that you have wine, $\bar{W}$ that you don't have wine, P that you had wine the previous day, and $\bar{P}$ that you didn't have wine the previous day. Do your preferences over such streams satisfy separability and stationarity?

Let's briefly return to the problematic kind of time-inconsistency, manifested by the common desire for vice today and virtue tomorrow. What could explain this phenomenon?

Part of the explanation might be that our preferences have 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 for tomorrow.

We could represent different sources of value by different subvalue functions. We might, for example, have a subvalue function V_c that measures the extent to which a proposition satisfies you present cravings, and another subvalue function V_m that measures to what extent it matches your moral convictions. Your intrinsic utility function is some kind of aggregate of these components. Here, too, separability is plausible. If, for example, you think that one world is morally better than another, and the two worlds are equally good with respect to all your other motives (your cravings are equally satisfied in either, etc.), then you plausibly prefer the first world to the second. This suggests that different sources of intrinsic utility combine in an additive manner.

7.5 Harsanyi's "proof of utilitarianism"

The ordinalist movement posed a challenge not only to the MEU Principle, but also to utilitarianism in ethics. Utilitarianism is a combination of two claims. The first says that an act is right iff it brings about the best available state of the world. The second says that the "goodness" of a state is the sum of the utility of all people. Without a numerical (and not just ordinal) measure of personal utility, this second claim makes no sense. 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. If we want to know which of two states is better, we can still ask which of them people prefer. This 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 7.10 (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, which of A and B is better? How about A and C, and B and C?

In 1955, John Harsanyi proved a remarkable theorem that seemed to rescue, and

indeed vindicate, classical utilitarianism.

As a first step, Harsanyi adopts von Neumann's response to the ordinalist challenge. He assumes that each individual has preferences not only among the relevant states, but also among lotteries involving the states, and that their preferences conform to the von Neumann and Morgenstern axioms. We can then represent their preferences by personal utility functions $U_1, \ldots, U_n$ (one for each individual) that are unique up to the choice of unit and zero.

Our goal is to derive a "social preference" relation between states that settles whether a state is overall better than another. Harsanyi assumes that this social preference relation can be extended to lotteries in a way that conforms to the von Neumann and Morgenstern axioms. It follows that social preference is also represented by a ("social") utility function U_s that is unique up to the choice of unit and zero.

Harsanyi now showed that if we add the Pareto condition (for both states and lotteries), then the individual and social preferences are represented by utility functions $U_1, \ldots, U_n$ and U_s in such a way that social utility is simply the sum of the individual utilities: for any state A,

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

Once we have allowed lotteries into the picture, the Pareto condition entails full-blown utilitarianism! How is this possible?

The Pareto condition implies that the social utility of any state is determined by the personal utility each individual assigns to the state. For suppose the social utility of some state A depends on an aspect of A that doesn't affect the personal utilities. Then there is an alternative B to A (that differs from A in this aspect) for which $U_s(B) \neq U_s(A)$ even though every individual assigns the same utility to A and B. This contradicts the Pareto condition.

So the only "attributes" of a state that are relevant to its social utility are its personal utility scores. We can represent a state by a list of numbers $\langle u_1, \ldots, u_n \rangle$, each of which specifies how desirable the state is for a particular individual.

Most non-utilitarians would disagree on this point. They would hold that even if everyone is indifferent between two states A and B, A might still be worse than B, if it involves gratuitous human rights violations, animal suffering, sin, or whatever.

The really surprising part of Harsanyi's theorem is that the social utility of a state

is simply the sum of its personal utility scores $u_1 + \ldots + u_n$. This tells us that social preference is separable across the personal utilities, and that each personal utility (each attribute) simply contributes its value to social utility. How does this come about? Couldn't an even distribution $\langle 10, 10, 10, 10, \ldots \rangle$ be better than an uneven distribution $\langle 0, 20, 0, 20, \ldots \rangle$? Relatedly, couldn't personal utility have "declining social value", so that adding 1 unit of personal utility to an individual whose utility is already at 1000 contributes less to social utility than adding 1 unit to an individual who stands at 0?

These possibilities are ruled out by three assumptions that look harmless in isolation, but have great power when combined.

One is the assumption that the Pareto condition holds for both lotteries and states. This implies that if every individual is indifferent between some lottery L and some state A, then the social preference relation is indifferent between L and A.

The second assumption is that each individual evaluates lotteries by their expected (personal) utility. Let L be a fair lottery between $\langle 0, 20, 0, 20, \ldots \rangle$ and $\langle 20, 0, 20, 0 \ldots \rangle$. The expected personal utility for each individual is 10. If everyone evaluates the lottery by its expected personal utility, then everyone is indifferent between L and $\langle 10, 10, 10, 10, \ldots \rangle$. By the first assumption, it follows that the social preference order is indifferent between L and $\langle 10, 10, 10, 10, \ldots \rangle$.

Finally, we have assumed that the social preference order ranks lotteries by their expected social utility. Assuming that the number of individuals is even, the states $\langle 20, 0, 20, 0, \ldots \rangle$ and $\langle 0, 20, 0, 20, \ldots \rangle$ plausibly have the same social utility. It follows that the social preference order is indifferent between either of these states and L. (If A and B have equal utility, then the expected utility of a lottery between A and B must equal the utility of A and B.) But we've just seen that the social preference order is indifferent between L and $\langle 10, 10, 10, 10, \ldots \rangle$. It follows that $\langle 0, 20, 0, 20, \ldots \rangle$ and $\langle 10, 10, 10, 10, \ldots \rangle$ have equal social utility.

If we think that even distributions of utility are better than uneven distributions, we have to reject at least one of the three assumptions. If we also accept that the right way to evaluate lotteries is by expected utility, it looks like the first assumption has to go. L is worse than $\langle 10, 10, 10, 10, \ldots \rangle$ even though each individual is indifferent between the two.

But should we accept that the right way to evaluate lotteries is by expected utility? This is the question to which we turn next.

Essay Question 7.1

Do you think time consistency is a requirement of rationality? Can you explain why, or why not?

Sources and Further Reading

The topic of this chapter is rarely discussed in mainstream philosophy, although its importance is occasionally recognized. See, for example, Philip Pettit, "Decision Theory and Folk Psychology" (1991). In economics, our topic is commonly known as "multi-attribute utility theory". Ralph L. Keeney and Howard Raiffa, *Decisions with Multiple Objectives* (1976/1993) is a classical, and very detailed, exposition. Paul Weirich, *Decision Space* (2001) explores the area from a more philosophical angle. The theorem by Debreu that I've referred to is from his 1960 article "Topological methods in cardinal utility". More results along the same line are surveyed in David Krantz et al., *Foundations of Measurement, Vol. I: Additive and Polynomial Representations* (1971).

For an in-depth discussion of preferences over time streams, including relevant empirical results, see Shane Frederick, George Loewenstein, and Ted O'Donoughue, "Time Discounting and Time Preference: A Critical Review" (2002).

A simple proof of Harsanyi's proof of utilitarianism is given in Michael D. Resnik, *Choices* (1987, pp. 197-200). For a sympathetic philosophical evaluation, see John Broome, "General and Personal Good: Harsanyi's Contribution to the Theory of Value" (2015).

8 Why MEU?

8.1 Arguments for the MEU Principle

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 gave a simple initial argument for the MEU Principle. An adequate decision rule, I said, 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.

In chapter 5, we looked at the internal structure of utility. I didn't mention it at the time, but the account we developed can be used to support the MEU Principle.

Consider a schematic decision matrix with n states $S_1, \ldots, S_n$. The expected utility of an act A is

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

In an adequate decision matrix, any act A in conjunction with any state S_i should determine the relevant outcome O_i, so that $S_i \wedge A$ entails O_i. Since outcomes have uniform utility, it follows that $\mathrm{U}(A \wedge S_i) = \mathrm{U}(O_i)$, for all i. Thus

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

In an adequate decision matrix, the states are independent of the acts. This suggests that $\mathrm{Cr}(S_i/A) = \mathrm{Cr}(S_i)$. So

$$\mathrm{EU}(A) = \mathrm{U}(A \wedge S_i) \cdot \mathrm{Cr}(S_1/A) + \ldots + \mathrm{U}(A \wedge S_n) \cdot \mathrm{Cr}(S_n/A).$$

In section 5.3, I mentioned a "partition formulation" of Jeffrey's axiom. This says

that for any proposition A and partition $S_1, \ldots, S_n$,

$$\mathrm{U}(A) = \mathrm{U}(A \wedge S_1) \cdot \mathrm{Cr}(S_1/A) + \ldots + \mathrm{U}(A \wedge S_n) \cdot \mathrm{Cr}(S_n/A).$$

Since the states in a decision matrix form a partition, it follows that $\mathrm{EU}(A) = \mathrm{U}(A)$: the *expected utility* of an act equals its *utility*.

It might seem strange to speak of an act's utility. When we use the MEU Principle, we assign *utilities* to outcomes and *expected utilities* to acts. We never talk about the utility of an act. In the terminology of chapter 5, each outcome is a "concern", as it settles everything the agent cares about. The theory of utility that we developed in chapter 5 allows us to extend an agent's "intrinsic" utility function for concerns to other propositions. In particular, we can talk about the utility of propositions that specify an act.

An act's utility measures how strongly the agent desires to perform the act. Assuming the theory of utility from chapter 5, the MEU principle reduces to the seemingly innocuous claim that rational agents choose an act that they desire to perform at least as strongly as any alternative. (We are going to challenge this seemingly innocuous claim in chapter 9.)

In chapter 6, we met yet another argument for the MEU Principle. The argument began with an idea about how to measure (or define) an agent's intrinsic utility function. The idea was to look at the agent's preferences between outcomes and lotteries. Assuming that the agent always chooses a most preferred option, von Neumann's construction of utility entails that an agent obeys the MEU Principle (in choices between lotteries) iff their preferences satisfy certain "axioms": Transitivity, Completeness, Continuity, Independence, and Reduction.

To complete this argument for the MEU Principle (for choices between lotteries), we would need to explain why the axioms should be considered requirements of rationality. Why should rational preferences satisfy Transitivity, Completeness, Continuity, Independence, and Reduction?

Here is an attractive answer: if an agent violates these axioms, then they will make patently bad choices in certain multi-stage decision problems.

To illustrate, suppose your preferences violate the Transitivity axiom. You prefer A to B, B to C, but C to A. Your preferences form a cycle. Whichever of A, B or C you have, you would prefer to have one of the others. If you are willing to pay a small amount to get the preferred option, it looks like I could exploit you in a kind

of multi-stage Dutch Book.

Concretely, let's assume you start out with C. Since you prefer B to C, you should be willing to pay an insignificant amount (say, 1p) if I let you swap C for B. Once you have B, I let you swap B for A in exchange for another penny. You should be happy to do that, given that you prefer A to B. Finally, I let you swap A for C, again in exchange for 1p. You should accept, as you prefer C to A. You are back where you started, with C, and I have gained three pence. We could start over, letting you swap C for B for A for C until I have emptied your wallet.

This kind of argument is called a **money-pump argument** (for obvious reasons). It's worth spelling out in more detail. In its present form, the argument has a serious flaw.

8.2 Money pumps and sequential choice

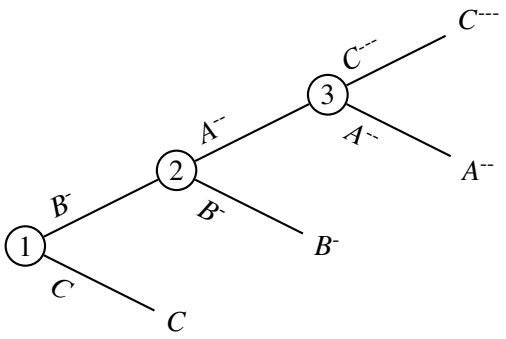
We are looking at an agent with cyclical preferences:

$$A \succ B$$
$$B \succ C$$
$$C \succ A$$

We imagine presenting this agent ("you") with a sequence of choices. A decision problem with more than one choice is called a **sequential decision problem**. The branch of decision theory that studies sequential decision problems is called **sequential decision theory** or **dynamic decision theory**. Our money-pump argument invites us to take a brief look into this area.

We have assumed that you start with C. At the first choice point in our money-pump scenario, you can either keep C or exchange C for B, at a small cost. Let B^- express B with the added small cost: $B^- = B\wedge$-1p. So your first choice is between C and B^-. If you choose B^-, you get the option to pay another penny to swap B for A. If you accept, you are left with $A^{--} = A\wedge$-2p. You are then offered a third choice, in which you can stick with A^{--} or end up with $C^{---} = C\wedge$-3p.

We can picture the whole sequential decision problem in a tree diagram, called an **extensive form representation**.



The circled nodes are choice points. What path through this tree would you take?

Above, I assumed that you would choose B^- at node 1. My reasoning was that you prefer B to C, and we take for granted that the preference is strong enough that you also prefer B^- to C. For analogous reasons, I assumed that you would choose A^{--} at node 2 (because you prefer A to B), and C^{---} at node 3 (because you prefer C to A). You end up with $C^{---} = C\wedge$-3p, even though you could have gotten C at no cost by "turning right" at the first node.

But would you really make these choices?

Look again at node 1. Superficially, you are here offered a choice between C and B^-. But if you "choose B^-" you aren't actually getting B^- unless you "turn right" at node 2. If you turn left at node 2 and again at node 3, as we assumed you will, then "choosing B^-" at node 1 actually means getting C^{---}. And C^{---} is worse than C. If you can foresee that you will turn left at nodes 2 and 3, then you will *not* turn left at node 1.

The flaw in my argument is that I have ignored any information you might have about your predicament and about what you might do at later stages in the scenario. We have adopted what is called a **myopic** approach to sequential choice. The myopic approach treats each choice as if it were the only decision the agent ever faces, ignoring any downstream consequences. We shouldn't be myopic. An adequate evaluation of the agent's options should take into account what the agent is likely to do later. This approach to sequential choice is called **sophisticated**.

To investigate our decision problem from a sophisticated perspective, we need to say what you know about your situation. Let's assume that you are fully informed about the sequential decision problem. Let's also assume that you have perfect knowledge of your preferences, so that you can figure out what you will do at any future choice point.

What you should do at node 1 now depends on what you might do at node 2, which similarly depends on what you might do at node 3. But if there are no relevant choices after node 3 then we can figure out what you would do here. The choice at node 3 is between A^{--} and C^{---}. Since you prefer C to A, it is plausible that you will choose C^{---}.

With this information in hand, we can return to node 2. Your choice at node 2 is effectively between C^{---} (*via* node 3) and B^{-}. You prefer B to C. So we can expect you to choose B^{-} at node 2.

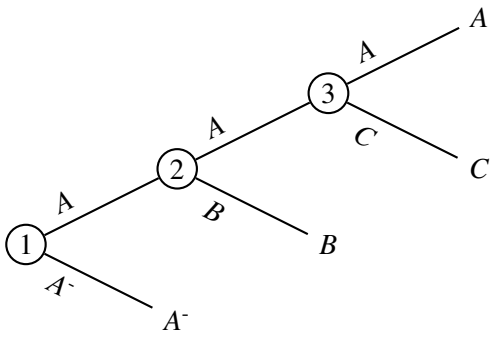
Now return to node 1. Given what we have just figured out, the choice at node 1 is effectively between C and B^{-}. You prefer B (and B^{-}) to C. We may therefore expect you to choose B^{-} at node 1. You will "turn left" at node 1 and right at node 2.

This kind of reasoning is called **backward induction**. We'll meet it again in section 10.5, where we will see that it is not as harmless at it might appear.

Exercise 8.1 †††

Draw a decision matrix (without utilities) for your choice at node 1.

The money pump argument from the previous section doesn't work – at least not if you know about my plot. But this can be fixed. In the following sequential decision problem, an agent who prefers A to B to C to A would trade A for A^{-} at node 1, assuming they know about the scenario and their preferences. They would make a guaranteed and avoidable loss of 1 penny.



Exercise 8.2 ††

Explain by backward induction why "you" (the agent with cyclical preferences) would choose A^- at node 1.

Exercise 8.3 †

Which choices would you make at which nodes if your preferences were transitive, so that $A \succ B$, $B \succ C$, and $A \succ C$?

The real point is, of course, not about money. The point is that cyclical preferences effectively lead to the choice of a dominated strategy. You could have gotten A, by "turning left" at each node. Due to your cyclical preferences, you end up with a strictly worse outcome A^-.

We have assumed that you prefer A to B, B to C, and C to A. Not all violations of Transitivity involve cycles of this kind. Instead of preferring C to A, you could be indifferent between C and A. You could also have no attitude at all about the comparison between A and C, violating both Transitivity and Completeness. These preferences, too, can be shown to support the choice of a dominated strategy. The same is true, more generally, for (almost) all preferences that violate the von Neumann and Morgenstern axioms.

8.3 The long run

I want to look at one more argument for the MEU Principle. This one turns on a connection between probability and relative frequency.

Suppose you repeatedly toss a fair coin, keeping track of the number of heads and tails. You will find that over time, the proportion of heads 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. 60% heads and 40% tails wouldn't be unusual. The (objective) 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 rules of probability entail that if there is a sequence of "trials" $T_1, T_2, T_3 \ldots$ in which the same outcomes (like heads and tails) can occur with the same probabilities, 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 **(weak) law of large numbers**. Loosely speaking: in the long run, probabilities turn into proportions.

How is this relevant to the MEU Principle? 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%. 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 an *average payoff*. In this kind of scenario, the MEU Principle effectively says that you should prefer acts with greater average utility (and therefore greater total utility) over acts with lower average (and total) utility. If you face the same decision problem over and over, then you are almost certain to achieve greater total utility if you follow the MEU Principle than if you follow any other rule.

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. Fortunately, we saw in the coin flip example that the convergence of proportions to probabilities tends to be quick. It does not take millions of tosses until the percentage of heads is almost certain to exceed 40%.

As it stands, the long-run argument still assumes that the same decision problem is faced over and over. But we can weaken this assumption. Suppose you face a sequence of decision problems that may involve different outcomes, different states, and different probabilities. One can show that if the states in these problems are probabilistically independent, and the relevant probabilities and utilities are not too extreme, then over time, maximizing expected utility is likely to maximize average (and total) utility.

From all this, you might expect that professional gamblers and investors generally put their money on the options with greatest expected payoff, since this 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 an investment in a startup that tries to find a cure for snoring. If the startup

succeeds, your investment will pay back tenfold. If the startup fails, 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 this gamble. If you repeatedly face this 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 illustrates that by maximizing expected utility you don't always make it likely that you will maximize average or total utility in the long run. More importantly, the example 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 the riskiness of their options?

Exercise 8.4 ††

Every year, an investor is given £100,000, which she can either invest 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?

(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?

8.4 Risk aversion

Many people are risk averse, at least for certain kinds of choices. They prefer situations with a predictable outcome over highly unpredictable situations. This does not seem irrational. Does it pose a threat to the MEU Principle?

A standard way to measure risk aversion involves lotteries. Consider a lottery with an 80% chance of £0 and a 20% chance of £1000. The 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?

We can. We could, for example, assume that the difference in utility between £1000 and £100 is, for this agent, 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.

This is how economists model risk aversion. They assume that for risk averse agents, utility is a "concave function of money", meaning that the amount of utility that an extra £100 would add to an outcome of £1000 is less than the amount of utility the same £100 would add to a lesser outcome of, say, £100. We have already encountered this phenomenon in chapter 5, where we saw that 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 valuable the more money you already have has nothing to do with risk. Money could have declining marginal utility even for an agent who loves the thrill of risky options. Conversely, an agent might value every penny as much as the previous one, but shy away from risks.

No doubt some actions that appear to display risk aversion (say, among professional gamblers) are really explained by the declining marginal utility of money. But many people prefer predictable situations in a way that can't be explained along these lines. The following example is due to Maurice Allais,

Example 8.1 (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 (0.8)	Green (0.19)	Blue (0.01)
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 (0.8)	Green (0.19)	Blue (0.01)
C	£1000	£1000	£1000
D	£1000	£1200	£0

If you choose C in the second choice, you get £1000 for sure. If you choose D, you get either £1000 (most likely) or £0 (least likely) or £1200. If you're risk averse, it makes sense to take the sure £1000.

In the first choice, the most likely outcome is £0 no matter what you do. It may seem reasonable to take the 19% chance of getting £1200 (by choosing B) rather than the 20% chance of getting £1000 (by choosing A).

Many people, when confronted with Allais's puzzle, seem to reason in this way. They prefer C to D and B to A. These preferences can't be explained by the declining marginal utility of money. Indeed, there is no way of assigning utilities to monetary payoffs that makes a preference of C over D and B over A conform to the MEU Principle. If you have the risk averse preferences, you appear to violate the MEU Principle.

Exercise 8.5 †††

The preference for C over D and B over A appears to violate the Independence axiom of von Neumann and Morgenstern. Explain. (The axiom states that, for any A, B, C, if $A \succsim B$, 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 $L_1 \succsim L_2$. You can assume Completeness.)

Some say that the kind of risk aversion that is manifested by a preference of B

over A and C over D is irrational. Rational agents, they say, can't prefer predictable situations over unpredictable situations. This might be OK if our topic were a special kind of "economic rationality". But it's not OK if we're interested in a general model of how coherent beliefs and desires relate to choice. There is nothing incoherent about a desire for predictability.

The following scenario, presented as a counterexample to the MEU Principle by Mark J. Machina, reinforces this verdict.

Example 8.2

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 the coin decide (C)	Abbie gets treat	Ben gets treat

The mother's preferences are $C \succ A$, $C \succ B$, $B \succ A$.

As in Allais's Paradox, there is no way of assigning utilities to the outcomes in the decision matrix in example 8.2 that makes the mother's preferences conform to the MEU Principle. Yet these preferences are surely not irrational. The mother prefers C because it is the most fair of the three options. It would be absurd to claim that rational agents cannot value fairness.

8.5 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 everything that matters to the agent.

Consider the bottom right cell of the second matrix in example 8.1. What will happen if you choose D and the blue ball is drawn? You get £0. But you might also feel frustrated about your bad luck: there was a 99% chance of getting at least £1000, and you got nothing! You probably don't like feeling frustrated. If so, the feeling should be included in the outcome. The outcome in the bottom right cell of the second matrix should say something like '£0 and considerable frustration'.

By contrast, consider the bottom right cell in the first matrix. If you choose B and the blue ball is drawn, you get £0. The chance of getting £0 was 81%, so you'll be much less frustrated about your bad luck. The outcome in that cell might say something like '£0 and a little frustration'. With these changes, the preference for B over A and 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 $\mathrm{EU}(B) > \mathrm{EU}(A)$ and $\mathrm{EU}(C) > \mathrm{EU}(D)$.

Do these changes reflect the values of a risk averse agent? Arguably not. Just as (genuine) risk aversion is not the same as declining marginal utility of money, it is not the same as fear of frustration. Imagine you face Allais's Paradox towards the end of your life. The ball will be drawn after your death, and the money will go to your children. You will not be around to experience frustration or regret. Nor might your children, if the whole process is kept secret from them. But if you like predictable outcomes, you might still prefer B to A and C to D.

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 (predictable) 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. The outcome should say something like '£0 and incurred avoidable risk'.

The outcome in the bottom right cell of the first matrix does not have the second attribute, that you have incurred an avoidable risk. There is no safe alternative in the

first matrix. We can once again distinguish the two outcomes, and reconcile your preferences with the MEU Principle.

Exercise 8.7 †

If you care about predictability and risk, then we should also distinguish the outcomes in all other cells of the matrix. Can you explain how?

Exercise 8.8 †

Redescribe the outcomes in example 8.2 so that the mother's preferences conform to the MEU Principle.

When social scientists discuss the MEU Principle, they generally assume that utility is assigned to material goods (as I mentioned in section 5.2). On this approach, an outcome in a decision matrix can only specify who owns which goods. Agents who care about frustration, predictability, or fairness are said to violate the MEU Principle.

There are reasons for this restricted conception of utility. Assuming that consumers maximise expected utility, in the restricted sense, and that material goods have declining marginal utility, one can derive various "laws" of microeconomics, such as the "law of demand". Even if people don't actually maximise expected utility, in the restricted sense, their behaviour as consumers might approximate what the economics version of our model predicts to make the model theoretically useful.

But our goal is not to derive the laws of microeconomics from substantive assumptions about what people ultimately care about. Our goal is to develop a general model of belief, desire, and rational choice. In this context, we don't want to put unnecessary and unrealistic constraints on what agents might desire. We want to allow for agents who care about frustration, predictability, fairness, and all sorts of other things.

Authors in the economics tradition sometimes consider models in which an agent's choices are assumed to be determined by their desire towards material goods, as reflected in their utility function, as well as their desire towards a specific further attribute – anticipated regret, for example, or riskiness. The MEU Principle is then revised to make room for the further parameter besides the agent's credence function

and the utility function for material goods. But this approach clearly doesn't generalise well. We have instead followed a popular tradition in philosophy that puts no substantive constraints at all on the objects of utility.

I should emphasize that these two approaches are not necessarily in tension. We are simply engaged in different projects.

A common objection to our unrestrictive conception of utility is that it seems to render the MEU Principle vacuous. In the economics interpretation, the MEU Principle predicts that rational agents don't choose B over A and C over D in Allais's Paradox. It also predicts that rational agents never toss a coin to decide who gets a treat. Our MEU Principle makes no such predictions. Indeed, for any pattern of behaviour, we can imagine that the agent has a basic desire to display just that behaviour. Displaying the behaviour then evidently maximizes expected utility. No behaviour whatsoever is, all by itself, ruled out by our MEU Principle.

This isn't necessarily a problem – not even for a descriptive understanding of the principle. Many respectable scientific theories are unfalsifiable *in isolation*. Scientific hypotheses can generally only be tested in conjunction with a whole range of background assumptions.

The same is true for the MEU Principle, understood as a descriptive hypothesis about human behaviour. *Given* some assumptions about an agent's beliefs and desires, we can easily find that their choices do not conform to the MEU Principle. And we often have good evidence about the relevant beliefs and desires. It is 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.

I said that any pattern of behaviour is compatible with the MEU Principle. Didn't von Neumann and Morgenstern prove that an agent maximizes expected utility in choices between lotteries *if and only if* their preferences (and therefore, one might think, their choice dispositions) satisfy some non-trivial conditions – Transitivity, Continuity, Independence, etc.?

Not quite. The proof of this result assumes that the agent's (intrinsic) utilities are determined by von Neumann's method. And here we reach a genuine downside to our approach: it breaks von Neumann's method.

Suppose, for example, we want to determine the intrinsic utility function for the mother in example 8.2. Let a and b be the outcomes of directly giving the treat to Abby and to Ben, respectively. If the mother cares about fairness, then one relevant aspect of both a and b is that the treat is not allocated through a chance process.

By von Neumann's method, we should now ask whether the mother prefers some other outcome c to a *lottery L 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. L entails that one of a and b comes about, and it also entails that neither of them come about. We can hardly assume that the mother has interesting views about how L compares to c.

In general, if we allow agents to care about arbitrary aspects of outcomes, then we can't assume that any lottery between outcomes is logically possible. Either the Completeness axiom or most of the other axioms become highly implausible.

This is a genuine cost. We lose a popular approach to defining utility, and a popular argument for the MEU Principle.

Exercise 8.9 †††

The money-pump argument for Transitivity from section 8.2 also makes substantive assumptions about what the agent ("you") ultimately cares about. Explain.

Similar problems arise for other attempts to measure utility in terms of preference, and to justify the MEU Principle. The popular theory of Leonard Savage, for example, also assumes that an agent's utility function pertains to a restricted set of "outcomes" that are logically independent of the "states" to which credences are assigned.

Ramsey's approach, however, might still work. Remember that instead of lotteries, Ramsey uses gambles of the form 'a if N, b if $\neg N$', where N is some proposition the agent doesn't care about and a and b are among the agent's concerns. If we understand such a gamble as a possible act that leads to a if N and otherwise to b, then the gamble may become logically impossible – if, for example, a entails that no such act is performed. But we don't have to interpret gambles as hypothetical acts. A gamble could simply be a certain kind of conditional proposition.

A clear example of a preference-based approach that imposes no substantive constraints on basic desires was developed by Ethan Bolker and Richard Jeffrey in the 1960s. Where von Neumann uses lotteries and Ramsey gambles, Bolker and Jeffrey use unspecific propositions. If a and b are two concerns, then the disjunction $a \vee b$ behaves somewhat like a lottery that "leads to" (i.e., amounts to) a with some prob-

ability (credence) and to b with another. As long as a and b are consistent, $a \vee b$ is guaranteed to be consistent as well.

Normally, the aim of a preference-based approach is to show that if an agent's preferences satisfy some plausible conditions ("axioms"), then the preferences can be represented by a utility function U, perhaps together with a credence function Cr, relative to which the agent ranks the things over which the preferences are defined by expected utility. Jeffrey and Bolker's preference relation is defined over arbitrary propositions. It isn't clear how we should understand the "expected utility" of, say, a disjunction $a \vee b$. But we've seen above, in section 8.1, that Jeffrey's concept of utility, which we have adopted since chapter 5, can be seen to generalise the concept of expected utility. Jeffrey and Bolker show that if an agent's preferences satisfy certain axioms, then the preferences can be represented by a utility function U and a credence function Cr relative to which the agent ranks propositions in line with Jeffrey's axiom.

So we might still be able to derive utility from preference – although the relevant preferences, relating arbitrary propositions, are even further removed from choice dispositions than in von Neumann's or Ramsey's construction.

Exercise 8.10 †††

Imagine you have an anti-rational streak: one of your basic desires is to *not maximise expected utility*. For simplicity, suppose your only other basic desire is to be free from pain, and it is weaker than your desire to not maximize expected utility. You wonder whether to bang your head against the wall. What does the MEU Principle say you should do?

Essay Question 8.1

In section 7.5, we looked at Harsanyi's argument for utilitarianism. The argument involves lotteries, and seems to rely on von Neumann's construction of utility. This suggests that the argument rests on implicit assumptions about what each individual may care about. Evaluate the prospects of trying to resist the argument on these grounds.

Sources and Further Reading

A useful survey of money-pump arguments for the von Neumann and Morgenstern axioms is Johan E. Gustafsson, *Money-Pump Arguments* (2022). Katie Steele, "Dynamic Decision Theory" (2018) briefly summarizes some of the philosophical controversy over these arguments.

I don't know any good literature on the long-run argument. I describe some moves towards generalising the argument beyond cases where the agent faces the same decision problem over and over at www.umsu.de/wo/2018/678.

For an intro to Allais's Paradox, see Philippe Mongin, "The Allais paradox: What it became, what it really was, what it now suggests to us" (2019). The example of the mother and the treat is from Mark J. Machina, "Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty" (1989).

That risk aversion should be handled by including risk as an "attribute" of outcomes is defended, for example, in Paul Weirich, "Expected Utility and Risk" (1986). For arguments against our liberal approach to utility, see Jean Baccelli and Philippe Mongin, "Can redescriptions of outcomes salvage the axioms of decision theory?" (2021) and chapter 4 of Lara Buchak, *Risk and Rationality* (2013).

The Jeffrey-Bolker construction is described in chapter 9 of Richard Jeffrey, *The Logic of Decision* (1965/83). Unless the agent's utilities are unbounded, Jeffrey and Bolker actually don't manage to secure a unique representation. On this issue, see, for example, James Joyce, "Why we still need the logic of decision" (2000).

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 with 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.

In an adequate decision matrix, the propositions we choose as the states must be independent of the acts. The need for this was illustrated in exercise 1.3. Here we looked at a student who wonders whether to study for an exam. The student drew up the following matrix and found, to her delight, that not studying is the dominant option.

	Will Pass	Won't Pass
Study	Pass & No Fun	Fail & No Fun
Don't Study	Pass & Fun	Fail & Fun

This is not an adequate matrix, unless the student is sure that studying would have no effect on the chance of passing. The states aren't independent of the acts.

What exactly does independence require? There are at least three notions of independence. A proposition A is **logically independent** of B 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 is **probabilistically independent** of B relative to some probability function Cr if $\mathrm{Cr}(A/B) = \mathrm{Cr}(A)$. (See section 2.4.) A is **causally independent** of B if whether or not B is true has no causal influence on whether A is true.

Exercise 9.1 †

In which of the three senses are the states in the student's decision matrix ('Will Pass', 'Won't Pass') independent of the acts, assuming that studying is known to increase the chance of passing?

When we require that the states in a decision matrix 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 accept 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

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

But EDT and CDT disagree on how the states should be construed. Each camp accuses the other of making a similar mistake as the student in exercise 1.3. If we require states to be probabilistically independent of the acts, the equation defines **evidential expected utility** ($\mathrm{EU_e}$); if we require causal independence, it defines **causal expected utility** ($\mathrm{EU_c}$).

Before we look at examples where EU_e and EU_c come apart, I want to mention three advantages of the evidential approach.

First, probabilistic independence is much better understood than causal independence. Provided $\mathrm{Cr}(B) > 0$, probabilistic independence between A and B simply means that $\mathrm{Cr}(A) = \mathrm{Cr}(A \wedge B)/\mathrm{Cr}(B)$. By contrast, our concept of causality or causal influence is often thought to be ill-defined and problematic. Bertrand Russell, for example, argued that "the word 'cause' is so inextricably bound up with misleading associations as to make its complete extrusion from the philosophical vocabulary desirable." 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 is supported by an argument I gave in section

8.1: assuming the theory of utility from section 5.3, one can show an act's evidential expected utility equals its utility. It will be useful to go over the argument again.

In section 5.3, we saw that Jeffrey's axiom is equivalent to the following schema, which I called the "partition formulation" of Jeffrey's axiom: for any partition $S_1, \ldots, S_n$ and proposition A with $\mathrm{Cr}(A) > 0$,

$$\mathrm{U}(A) = \mathrm{U}(A \wedge S_1) \cdot \mathrm{Cr}(S_1/A) + \ldots + \mathrm{U}(A \wedge S_n) \cdot \mathrm{Cr}(S_n/A). \tag{J1}$$

As a special case, assume that $S_1, \ldots, S_n$ is a partition that is fine-grained enough so that every conjunction of A with a member of the partition settles everything the agent ultimately cares about. That is, for each S_i in the partition, $A \wedge S_i$ entails one of the agent's concerns. Let O_i be the concern entailed by $A \wedge S_i$. We then have $\mathrm{U}(A \wedge S_i) = \mathrm{U}(O_i)$. Plugging this into (J1), we get

$$\mathrm{U}(A) = \mathrm{U}(O_1)\mathrm{Cr}(S_1/A) + \ldots + \mathrm{U}(O_n)\mathrm{Cr}(S_n/A). \tag{J2}$$

Now suppose we have drawn up a decision matrix that conforms to the evidentialist requirement that the states are probabilistically independent of the acts. Let $S_1, \ldots, S_n$ be the states in this matrix. The evidential expected utility of an act A is defined as

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

Each conjunction of an act A with one of the states S_i settles everything the agent cares about. So equation (J2) applies. Moreover, the states are probabilistically independent of the acts: $\mathrm{Cr}(S_i/A) = \mathrm{Cr}(S_i)$, for all i. It follows that $\mathrm{EU_e}(A) = \mathrm{U}(A)$.

The MEU Principle, as understood by EDT, says that rational agents choose acts that are at least as desirable as the available alternatives. Friends of CDT have to deny this. They hold that rational agents sometimes choose undesirable acts even though they could have chosen a more desirable alternative. On the face of it, the EDT account looks more plausible.

A third advantage of EDT is that it allows computing expected utilities in a way that is often simpler and more intuitive than the method we've used so far.

We've seen that an act's evidential expected utility equals the act's utility, as determined by Jeffrey's axiom. We can therefore use (J1) or (J2) to compute $\mathrm{EU_e}$. These equations are correct even for partitions $S_1, \ldots, S_n$ whose members are *not* independent of the act A.

Return to the student's decision problem. The problem with her matrix is that the 'Will Pass' state is more likely if the student studies than if she doesn't study. Intuitively, we should give greater weight to 'Will Pass' when we evaluate the option 'Study' than when we evaluate 'Don't Study'.

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 matrix, 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*. She is only 20% confident that she will pass *if she doesn't study*, as indicated by '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 the relevant cells and add up the products. 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, our **new method** for computing expected utilities works as follows. As before, we need to 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). All we require is that each state in combination with each act settles everything the agent cares about. If an act A leads to outcomes $O_1, \ldots, O_n$ in states $S_1, \ldots, S_n$ respectively, then we compute the expected utility of A as

$$\text{EU}_\text{e}(A) = \text{U}(O_1) \cdot \text{Cr}(S_1/A) + \ldots + \text{U}(O_n) \cdot \text{Cr}(S_n/A).$$

The unconditional credences $\text{Cr}(S_i)$ in the old method have been replaced by conditional credences $\text{Cr}(S_i/A)$, to compensate for the fact that the states may not be independent of the acts.

When we compute an act's expected utility with this new method, we are effectively using (J2) to determine the act's utility, which we know equals the act's eviden-

tial expected utility. The expected utility determined by the new method is evidential expected utility.

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 evidential expected utilities by the old method.
(b) Use the new method to compute the (evidential) expected utilities.

We can go further. Let $O_1, \ldots, O_n$ be the possible outcomes of act A (or more generally, the concerns that are logically compatible with A). Any conjunction of O_i and A obviously entails one of the outcomes – namely O_i. We can therefore choose the outcomes themselves as the partition $S_1, \ldots, S_n$ in (J2). We get

$$\mathrm{U}(A) = \mathrm{U}(O_1)\mathrm{Cr}(O_1/A) + \ldots + \mathrm{U}(O_n)\mathrm{Cr}(O_n/A). \tag{J3}$$

This suggests yet another way of computing expected utilities. I'll call it the **state-free method**. When we use the state-free method, we only need to figure out all the outcomes $O_1, \ldots, O_n$ a given act might bring about. We then consider how likely each of these outcomes is on the supposition that the act is chosen, and take the sum of the products:

$$\mathrm{EU_e}(A) = \mathrm{U}(O_1) \cdot \mathrm{Cr}(O_1/A) + \ldots + \mathrm{U}(O_n) \cdot \mathrm{Cr}(O_n/A).$$

By (J3), the result is the act's utility, and therefore the act's evidential expected utility.

In practice, the new method and the state-free method are often simpler and more

intuitive than the old method.

Exercise 9.3 †

I offer you a choice between £10 for sure and a coin flip that would give you £20 on heads or £0 on tails. The coin will not be flipped if you take the first option. In cases like this, it is hard to find a suitable set of states. Use the state-free method to compute the expected utility for the two options, assuming your intrinsic utility equals monetary payoff.

Exercise 9.4 †††

When I derived the state-free method, I assumed that the different outcomes an act might bring about form a partition. Explain why this is not generally true, and why (J3) is correct nonetheless.

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 blue box and a transparent box. The transparent box contains £1000. You can't see what's in the blue box. You have two options. You can take *just the blue box* and keep whatever is inside. Alternatively, you can take *both boxes* and keep their content. Last night, a demon has scanned your brain, trying to predict what you will do. (You knew nothing about this at the time.) If she predicted that you would take both boxes, then she has put nothing in the blue box. If she predicted that you would take just the blue box, she has 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 and have no other relevant desires?

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 is adequate, according to CDT.

	£1,000,000 in blue box	£0 in blue box
Take only blue 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 blue 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 blue box in response to your choice. Her only superpower is predicting people's choices from their brain state in the previous night.

It is obvious from the decision matrix that taking both boxes maximizes causal expected utility. Two-boxing 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 the demon's prediction is correct. Your credence that there's a million in the blue box is 0.95 on the supposition that you one-box and 0.05 on the supposition that 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" from the previous section, the evidential expected utility of the two options then works out as follows ('1B' is one-boxing, '2B' is two-boxing):

$$\begin{aligned} \mathrm{EU_e}(1\mathrm{B}) &= \mathrm{U}(£1{,}000{,}000) \cdot \mathrm{Cr}(£1{,}000{,}000/1\mathrm{B}) + \mathrm{U}(£0) \cdot \mathrm{Cr}(£0/1\mathrm{B}) \\ &= 1,000,000 \cdot 0.95 + 0 \cdot 0.05 = 950,000. \\ \mathrm{EU_e}(2\mathrm{B}) &= \mathrm{U}(£1{,}001{,}000) \cdot \mathrm{Cr}(£1{,}001{,}000/2\mathrm{B}) + \mathrm{U}(£1000) \cdot \mathrm{Cr}(£1000/2\mathrm{B}) \\ &= 1,001,000 \cdot 0.05 + 1000 \cdot 0.95 = 51,000. \end{aligned}$$

One-boxing comes out better than two-boxing.

CDT says that you should two-box; EDT says you should one-box. Who is right? Philosophers have been debating the question for over 50 years, with no consensus in sight.

Some think one-boxing is obviously the better choice. 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. Most two-boxers walked away with a thousand. Wouldn't you rather be in the first group than in the second? It's your choice!

Practical rationality is all about satisfying your goals in the light of your beliefs. We have stipulated that the only goal in Newcomb's Problem is to get as much money as possible. It seems obvious that one-boxing is the better strategy for achieving this goal. One-boxing is the ticket to a million, two-boxing to a thousand.

Others think it equally obvious that you should two-box. If you take both boxes you are guaranteed to get £1000 more than whatever you'd get if you took just the blue box. Remember that the content of the boxes is settled. The blue box either contains a thousand or a million. One-boxing and two-boxing both give you the blue 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.

Here's another argument for two-boxing. Imagine you have a friend who helped the demon prepare the boxes. Your friend knows what's in the blue 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 blue 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 signal to take both boxes, so that you get at least the thousand. If the box contains a million, she will also signal 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 they ought to do *in light of their beliefs*. In the Miners Problem (example 1.1), the best choice in light of all the facts is to block whichever shaft the miners are in. Since you don't know where the miners are, you don't know which of your options is best in light of all the facts. You have to go by your limited information. The best choice

in light of your 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. You also know what you would decide to do if *you* knew what's in the blue box: You would (plausibly) take both boxes. EDT says that you should one-box even though you know that two-boxing is best in light of all the facts!

Exercise 9.5 ††

Imagine that before you make your choice, the demon threatens to reveal the contents of the blue box, unless you pay them £100,000. Explain why EDT says that you should pay.

What about the fact that one-boxers are generally richer than two-boxers? Doesn't this show that the one-boxers are doing something right? Not so, say those who advocate two-boxing. Compare: people who fly business class are generally richer than people who fly economy. Clearly this doesn't show that everyone should fly business class. Flying business class wouldn't *make* you rich. Similarly for one-boxing. All the one-boxers who got a million are rich not because they made a great choice but because they were given great options. They were put in front of a blue box containing a million and a transparent box containing a thousand. They were, in effect, given a choice was between £1,001,000 and £1,000,000. It's not a great achievement that they walked away with a million. All the two-boxers who got a mere thousand were effectively given a choice between £1000 and £0.

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 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 blue box (which you're certain to get) contains a million. That's the sense in which one-boxing is desirable. You should be delighted to learn that you are going to one-box. Two-boxing, by contrast, is bad news. It indicates that the blue box is empty. 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 blue box, then it got in there long before you made your choice.

This difference between EDT and CDT can show up in more realistic scenarios. Some versions of the Prisoner's Dilemma (example 1.3) are plausible candidates. Suppose you only care about your own prison term. We can then represent the Prisoner's Dilemma by the following matrix.

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

The "states" (your partner's choice) are causally independent of the acts. No matter what your partner does, confessing leads to a 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 $\mathrm{Cr}(\text{she confesses}/\text{you confess}) = \mathrm{Cr}(\text{she is silent}/\text{you are silent}) = 0.8$. 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 9.6 †

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 cancer rates are 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. Cancer, on that explanation, isn't caused by smoking, but 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 it then they would keep smoking. According to EDT, however, it seems that people should give up smoking even if they believed the tobacco industry's story.

Let's work through a toy model to see why. 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 (S) and getting cancer (C) are as follows:

$$\begin{aligned} \mathrm{U}(S \wedge \neg C) &= 1 \\ \mathrm{U}(\neg S \wedge \neg C) &= 0 \\ \mathrm{U}(S \wedge C) &= -9 \\ \mathrm{U}(\neg S \wedge C) &= -10 \end{aligned}$$

Suppose you are convinced by the tobacco industry's explanation: you are sure that smoking does not cause cancer, but that it indicates the presence of a cancer-causing gene. So $\mathrm{Cr}(C/S)$ is greater than $\mathrm{Cr}(C/\neg S)$. Let's say $\mathrm{Cr}(C/S) = 0.8$ and $\mathrm{Cr}(C/\neg S) = 0.2$. 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, you should stop smoking. Indeed, it should make no difference to you whether smoking causes cancer or merely indicates a predisposition for cancer. Either way, smoking is bad news.

This is not what the tobacco industry expected. And it does seem odd. You are sure that smoking will not bring about anything bad. On the contrary, smoking is guaranteed to make things better. At the same time, it would be evidence that you have a bad gene. By not smoking, you can suppress this 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 say, smoking actually won't be evidence for cancer: $\mathrm{Cr}(C/S)$ won't be

greater than $\mathrm{Cr}(C/\neg S)$. We have assumed that the gene causes smoking by causing a desire to smoke. But suppose you feel a strong desire to smoke. The 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. Once you've taken into account the information you get from the presence or absence of the desire, $\mathrm{Cr}(C/S) = \mathrm{Cr}(C/\neg S)$. And then EDT recommends smoking (in our fictional scenario).

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 9.7 †

You wonder whether to vote in a large election between two candidates A and B. You assign subvalue 100 to A winning and -100 to B winning. Voting would add a subvalue of -1, since it would cause you some inconvenience. You are confident that the election will be close, but almost sure (credence around 0.9999) that it won't come down to a single vote. You think you are typical for a certain group of A supporters: you estimate that around 1-2% of A's supporters will reach the same decision about whether to vote that you will reach, based on the same reasons. Explain, without computing anything, why CDT says that you shouldn't vote, but EDT says you probably should.

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 expected utility is the classical definition in terms of states, acts, and outcomes, where we now require that the states are *causally independent* of the acts. But one can also construct a version of CDT that looks more like EDT, and shares at least some of EDT's attractive features. The key to this construction is a point I briefly mentioned in section 2.4: that there are two ways of supposing a proposition.

Throughout the Second World War, Nazi Germany tried to develop nuclear weapons. Consider the hypothesis that these attempts succeeded in 1944. If we entertain the hypothesis as a **subjunctive** or **counterfactual** supposition, we wonder what *would have happened* if the attempts had succeeded. Knowing Hitler's character, it is likely that he would have used the weapons, possibly leading to an axis victory in the war.

In general, when we subjunctively suppose that an event took place, we try to figure out 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.

Things are different when we **indicatively suppose** that the Nazis had nuclear weapons in 1944. Here we hypothetically add the supposed proposition to our beliefs and revise the other beliefs in a minimal way to restore consistency. We know, for example, that Hitler didn't use nuclear weapons. Supposing that Germany had nuclear weapons, we infer 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. 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 A and B and logical combinations of these. Whether B would be the case on the supposition that A had been the case generally depends on the laws of nature and various particular facts besides A and B.

Now return to the "new method" for computing (evidential) 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 the states and acts are independent:

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

These are indicative conditional probabilities. If we use subjunctive conditional probabilities, we get a formula for causal expected utility:

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

Admittedly, it isn't obvious that this is equivalent to our original definition of $\mathrm{EU_c}$ in terms of "causally independent" states. To establish the equivalence, we

would have to say more about the relevant notion of causal independence and about subjunctive supposition.

There are, in fact, many different proposals on the market for how CDT should be spelled out. We have seen two. They may not be equivalent, but both are "causal" insofar as they involve broadly causal notions in the definition of expected utility.

The above formula for $\mathrm{EU_c}$ can be used with any partition $S_1, \ldots, S_n$ that is sufficiently fine-grained so that each conjunction $S_i \wedge A$ settles everything the agent cares about. As before, we can therefore use the outcome partition as $S_1, \ldots, S_n$ to get a state-free formula:

$$\mathrm{EU_c}(A) = \mathrm{U}(O_1) \cdot \mathrm{Cr}(O_1 /\!/ A) + \ldots + \mathrm{U}(O_n) \cdot \mathrm{Cr}(O_n /\!/ A).$$

To get a feeling for how this works, let's 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. To consider the causal expected utility of taking the box, we suppose, subjunctively, that you take the box. We ask: how much you would get if you were to take the box?

Answer: it depends on what's inside. In a world where the box contains £1,000,000, you would get £1,000,000 if you were to take the box. In a world where the box contains £0, you would get £0. Both possibilities have equal probability. So

$$\mathrm{Cr}(\text{£1,000,000 // Take box}) = 0.5$$
$$\mathrm{Cr}(\text{£0 // Take box}) = 0.5.$$

In general, if you have the option of taking a box that contains a certain amount of money, 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 inside. Any uncertainty about how much you would get boils down to uncertainty about how much is in the box.

Exercise 9.8 ††

Use the state-free method for computing causal expected utility to evaluate the two options in Newcomb's problem.

Exercise 9.9 †††

Consider the second 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 £120 in the red box and £130 in the green box. If she predicted that you would take the green box, she put £70 in the green box and £90 in the red box. If she predicted that you would take the transparent box, she put £100 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	£120	£90	£100
G	£130	£70	£100
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. But suppose you're inclined to take the red box. At this point, it is no longer rational to treat all three predictions as equally likely:

you should become confident that the demon has predicted R. And then R no longer maximizes expected utility. You should reconsider your choice.

Exercise 9.10 ††

Can you see where this process of deliberation will end? (Explain briefly.)

It is even possible 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 runs into Death, who looks surprised. "I am coming for you tomorrow", Death says. Terrified, the man buys a horse and rides all through the night to Aleppo, where he plans to hide in a hidden alley. As he enters the alley, he sees Death waiting for him. "I was surprised to see you yesterday 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 will be tomorrow. Like in Newcomb's Problem, let's assume the prediction is settled, and not (causally) affected by what you decide to do. But Death is a very good predictor. If you go to Aleppo, you can be confident that Death will wait for you there. 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?

Essay Question 9.1

What is the rational choice in Newcomb's Problem? Can you think of an argument for either side not mentioned in the text?

Sources and Further Reading

Newcomb's Problem was first discussed in Robert Nozick "Newcomb's Problem and Two Principles of Choice" (1969). That a better-informed friend would advise you to two-box is noted already by Nozick. That two-boxing is known to be better in light of all the facts is noted in Jack Spencer and Ian Wells, "Why Take Both Boxes?" (2017). That EDT recommends paying to not find out what's in the blue box is noted in Brian Skyrms, "Causal Decision Theory" (1982). For more on realistic Newcomb cases and the tickle defence, see chapter 4 of Arif Ahmed, *Evidence, Decision, and Causality* (2014).

The classical exposition of EDT is Richard Jeffrey, *The Logic of Decision* (1965/1983). Classical expositions of CDT include Allan Gibbard and William L. Harper, "Counterfactuals and two kinds of expected utility" (1978), David Lewis, "Causal Decision Theory" (1981), and James Joyce, *The Foundations of Causal Decision Theory* (1999).

"Death in Damascus" is discussed in the Gibbard & Harper paper. For more on the theme of section 9.5, start with Frank Arntzenius, "No Regrets, or: Edith Piaf Revamps Decision Theory" (2008).

10 Game Theory

10.1 Games

Game theory studies decision problems in which the outcome of an agent's choice depends on other agents' choices. Such problems are called **games**, and the agents **players**. The Prisoner's 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 they 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 we may be able to shed light on important social and political issues. The way we live and behave, as a society, is in many ways not ideal. We are depleting the Earth's resources. We are destabilising the climate. We are woefully underprepared for pandemics and other catastrophes. We buy goods from online retailers where most of the products are a scam. Corruption is rampant. The political system is broken. Dating is broken. And so on, and on. Why? Why don't we fix these problems? Is it because the current system benefits powerful actors who have us under their control? Game theory suggests an alternative possibility.

Remember the Prisoner's Dilemma. If you and your partner are rational and don't care about each other, you both confess and spend a long time in prison. Collectively, you could have achieved a much better outcome by remaining silent. Things are unnecessarily bad – you spend a long time in prison – not because a powerful third party stands to gain from your misery. The bad outcome is simply a result of your misaligned incentives.

This kind of situation is sadly common. Professional athletes, for example, have a strong incentive to use steroids, as long as the chance of being caught is low. Whether

or not their competitors do the same, using steroids provides an advantage. The outcome is that everyone uses steroids, even though everyone would prefer that no-one uses steroids. Structurally, the athletes' decision problem is the same as the Prisoner's Dilemma. Any decision problem with this structure is nowadays called a Prisoner's Dilemma, even if no prisoners are evolved.

Another famous example is the "tragedy of the commons". Fishermen have an incentive to catch as many fish as they 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's 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's Dilemma.

Exercise 10.1 †

How do criminal organisations like the Mafia ensure that its members remain silent when they are interrogated by the police? Draw the decision matrix for the scenario of the (original) Prisoner's Dilemma, but assuming that both players are members of the Mafia.

Another reason to study games is that a new set of conceptual tools and techniques become available if the states in a decision problem are other people's actions. In particular, we can often figure out which state obtains based on the other players' desires. In the original Prisoner's Dilemma, we know that if your partner is rational any only cares about their own prison term then they will confess.

Here is how game theorists would typically draw the matrix for the Prisoner's Dilemma, assuming you and your partner don't care about each other:

	Confess	Silent
Confess	-5, -5	0, -8
Silent	-8, 0	-1, -1

As before, the rows are the acts available to you. The columns are the acts available to your partner. We generally don't assign credences to the columns. The numbers in the cells represent the utility of the relevant outcome for you and your partner. We

don't describe the outcome itself any more, for lack of space. The first number in each cell is the utility for the row player (whom we'll call 'Row' and assume to be female); the second is the utility for the column player ('Column', male).

In game theory jargon, a **solution** to a game is a prediction of what each player is going to do, assuming that they are rational. The solution to the Prisoner's Dilemma is that each player confesses. Confessing dominates remaining silent. You should confess no matter what you think your partner will do.

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. C_2 dominates C_1. So if Row knows the utility that Column assigns to the outcomes, then she can figure out that Column will choose C_2. And so Row should choose R_2. The solution is R_2, C_2: Row chooses 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. 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.

To reach this conclusion, we need to assume more than that both players know each other's utilities. 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.

A common idealisation in game theory is that the players have **complete information** about the game, meaning that

(1) all players know the structure of the game, as displayed in the matrix;
(2) all players know that all players are rational;
(3) all players know that (1)–(3) are satisfied.

By applying to itself, the clause (3) 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**. (1)–(3) say that the structure of the game and the rationality of all participants are common knowledge.

Exercise 10.2 ††

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	2, 3
R_2	2, 1	3, 2	4, 1
R_3	3, 3	1, 1	4, 2

No option for either player is dominated by any other. Can we nonetheless figure out what Row and Column will choose?

Let's start with some trial and error. Take R_1, C_1. Could this be how the game is always played, 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 or C_3. What about R_1, C_2? If this is how the game has to played, then Row would know that Column plays C_2, and then she would be better off playing R_2. This kind of reasoning disqualifies all 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, he 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 their part of the equilibrium, given that the other players stick to their part.

There is a simple algorithm for finding Nash equilibria in two-player games. 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 4 in the first column, the 3 in the middle cell, and both 4s in the third column. Then do the same for 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, $\underline{3}$	2, $\underline{3}$
R_2	2, 1	$\underline{3}$, $\underline{2}$	$\underline{4}$, 1
R_3	3, $\underline{3}$	1, 1	$\underline{4}$, 2

Any cell in which both numbers are underlined is a Nash equilibrium.

A common assumption in game theory is that if a game has a unique Nash equilibrium, and assumptions (1)–(3) are satisfied, then the Nash equilibrium is the game's solution: each player will play their part of the equilibrium.

But this isn't obvious. Our trial-and-error reasoning from above shows that *if a game has a unique solution*, and assumptions (1)–(3) are satisfied, *then the solution is a Nash equilibrium*. The reason is that if a game has a unique solution, then (1)–(3)

entail that each player knows that the other will play their part of the solution. Each player plays their part of the solution with the full knowledge that the other player is playing their part. So the solution must be a Nash equilibrium.

It doesn't follow, however, that if a game has a unique Nash equilibrium, then this is the game's solution. Consider the following 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

There is a unique Nash equilibrium: R_3, C_2. If this is the guaranteed outcome under assumptions (1)–(3), then Row can be sure that Column will play C_2. But if Column plays C_2, then R_2 and R_3 are equally good for Row. So how can we be sure Row won't play R_2?

You might argue that if Row played R_2 and Column could predict her choice, then Column would play C_3, leading to a worse result for Row. But we're not assuming that Column can predict Row's choice. All we're assuming is (1)–(3).

A better argument in support of R_3, C_2 as the unique solution goes as follows. Suppose for reductio that Row could play either R_3 or R_2, and conditions (1)–(3) are satisfied. Then Column can't be sure that Row will play R_2. If Column gives equal credence to R_2 and R_3, then his best choice is C_3. And then Row should choose R_3, contradicting our assumption that Row can play R_2.

This argument is still a little shaky. Why would Column have to give equal credence to R_2 and R_3? Why couldn't Column be confident that Row will play R_3 and yet Row actually plays R_2?

We need more than (1)–(3) to ensure that a unique Nash equilibrium will always be played. We seem to need the further assumption that each player can replicate the other's process of deliberation – or at least the end point of the process.

One reason to think that this assumption might be satisfied is that the players seem to have the same evidence about the game. If the norms of rationality determine how, say, Column should figure out what he should do, based on his evidence and his goals, then Row – knowing Column's evidence, his utilities, and his rationality – can replicate Column's process of deliberation: she can figure out how Column will figure out what he should do.

Another, simpler, reason why each player may know about the other's deliberation is that they have played the same game before. In repeated plays, each player has direct evidence about how the other tends to play, from what they did on the previous iterations. If, in the above example, Row always plays R_2, then Column will start playing C_2. Seeing that Column plays C_2, Row should switch to R_1 or R_3. Eventually, we would expect them to end up in the Nash equilibrium R_3, C_2.

Exercise 10.3 †

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 10.4 ††

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 this 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 one 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 sum to the same number. Since utility scales don't have a fixed zero, we can re-scale the utilities so that the sum is zero. For this reason, games in which the players' preferences are opposed are called **zero-sum games**. Here is an example.

	C_1	C_2	C_3
R_1	1, -1	3, -3	1, -1
R_2	2, -2	-2, 2	-1, 1

There is a unique Nash equilibrium: R_1, C_3. Curiously, this equilibrium will be reached if each player follows the *maximin* rule that we've met in section 1.4. Maximin says to choose an option with the best worst-case result. In our example, the worst-case result of choosing R_1 (for Row) has utility 1; the worst-case result of R_2 is -2. Maximin therefore says that Row should choose R_1. For Column, it similarly recommends C_3.

This is not a coincidence. Every Nash equilibrium in every zero-sum game is supported by the maximin rule. For suppose that R_i, C_j is a Nash equilibrium in a (two-player) zero-sum game, but R_i isn't supported by the Maximin rule. Then there is an alternative R_k whose worst-case outcome is better for Row than the outcome of R_i, C_j. Then every possible outcome of R_k is better for Row than R_i, C_j. But if R_i, C_j is a Nash equilibrium, then R_k, C_j can't be better for Row than R_i, C_j.

One might argue that even though maximin is not a generally defensible decision rule, it makes sense in a zero-sum game with complete information. The idea would be that whatever option R_i Row chooses, she can be confident that Column will choose an option C_j that leads to the best outcome when combined with R_i. And the best outcome for Column is the worst outcome for Row. Like the argument for Nash equilibria in the previous section, however, this argument assumes that the players can replicate each other's reasoning.

Many games have more than one Nash equilibrium. The hypothesis that players usually end up in a Nash equilibrium then doesn't fully tell us what the players will do. Here is an 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

There are fur Nash equilibria. What will the players do? Should Row play R_1 or R_2? Should Column play C_2 or C_3? Well, it doesn't matter. The players 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 10.5 †††

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.

	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. What should you do in this kind of game?

A standard answer in game theory is that you should randomize. You should, say, toss a fair die and choose Rock on 1 or 2, Paper on 3 or 4, and Scissors on 5 or 6. Such a randomized choice is called a **mixed strategy**. We will write '[1/3 Rock, 1/3 Paper, 1/3 Scissors]' for the mixed strategy of playing Rock, Paper, or Scissors each with (objective) probability 1/3.

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). 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 finite 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 hate randomization and prefer losing in Rock–Paper–Scissors to randomizing, then the game has no Nash Equilibrium, not even among mixed strategies.

Exercise 10.6 ††

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 the players 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

There are two Nash equilibria (without randomization): both going to party A, and both going to party B. The first equilibrium is better, but our assumptions (1)–(3) appear to be compatible with either. One can imagine a scenario in which you and Bob are both confident that the other will go to party B. Going to B then maximises expected utility. One can also imagine a scenario in which you are confident that Bob will go to B and Bob is confident that you will go to A, so that you end up at different parties. If we don't assume that you can replicate each other's reasoning, all outcomes appear to be possible.

I say 'appear' because it isn't obvious what credences are rationally permitted in this situation. Could you be rationally confident that Bob will go to B, under conditions (1)–(3)? You can figure out that Bob will go to B iff he is more than 60%

confident that *you* will go to B. So the question is, can you be confident that Bob is more than 60% confident that you will go to B? Of course, Bob knows that you will go to B only if you are more than 60% confident that he will go to B. So the question is, can you be confident that Bob is more than 60% confident that you are more than 60% confident that he will go to B? And so on. There is nothing incoherent about this state of mind, in which you are confident that Bob will go to B. But we may wonder how you could have rationally arrived at this state.

Our assumptions (1)–(3) here give rise to an epistemological puzzle. If you have no further relevant evidence, how confident should you be that Bob will go 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.

Another tempting thought is that you must be sure that Bob will go to A. But why? What part of your evidence rules out scenarios in which he goes to B?

Exercise 10.7 ††

Suppose you know that Bob can replicate your reasoning. What does Evidential Decision Theory say you should do in the party situation (example 10.1)?

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 themselves. There are two Nash Equilibria in Chicken that don't involve randomization: 'Swerve, Straight' and 'Straight, Swerve'. As above, every 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 irrational 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. Bob will win.

There are rumours 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 10.8 †

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 10.9 †††

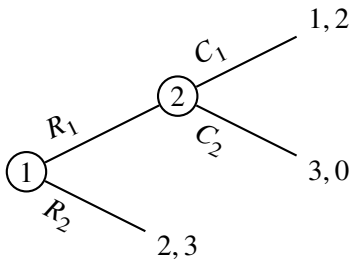
A third Nash equilibrium in Chicken involves randomization. Can you find it?

10.5 Games with several moves

So far, we have looked at games in which each player makes just one move, and no player knows about the others' moves ahead of their choice. Game theory also studies situations 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.

As in section 8.2, we can picture the relevant decision situations in a tree-like diagram (an "extensive form representation"). Below is a diagram for a game in which Row first has a choice between R_1 and R_2. If she chooses R_2, the game ends

with an outcome that has utility 2 for Row and 3 for Column. If Row chooses R_1, then Column gets a choice between C_1 and C_2. If he chooses C_2, Row gets utility 3 and Column 0; if Column chooses C_1, Row gets 1 and Column 2.



We can use backward induction to predict how the game is going to be played, assuming (1)–(3).

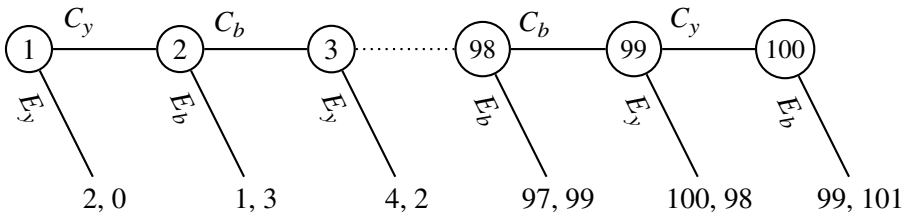
Consider node 2, where Column has a choice between outcome '3, 0' and outcome '1, 2'. The choice involves no relevant uncertainty, and Column prefers '1, 2' over '3, 0'. He can be expected to play C_1. Anticipating this, Row can figure out that playing R_1 at node 1 will lead to '1, 2'. R_2 instead leads to '2, 3'. This is better for Row. So Row will play R_2.

In the following example, backward induction leads to a more surprising result.

Example 10.3 (Centipede)

You and Bob are playing a game. The game starts with a pot containing £2. In round 1, you can decide 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), the pot is divided so that he gets £3 and you get £1. If he continues, the money in the pot increases by another £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.



Let's use backward induction to solve the game. At node 100, Bob doesn't have a choice. If you continue at node 99 (C_y), you will get £99 and Bob £101. If you end the game (E_y) at node 99, you will get £100. It is obviously better to end the game. Anticipating this, what should Bob do in round 98? If he ends the game (E_b), he'll get £99; if he continues (C_b), 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 that the game should be ended. In particular, you can anticipate in round 1 that Bob will end the game in round 2. So you should end the game in round 1. You will get £2 and Bob £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 an MEU perspective. It is clear what happens if you end the game: you'll get £2. But what would happen if you chose to continue? The argument from backward induction assumes that Bob would end the game. 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. And if Bob faces a choice in round 2, then he has just seen that you *did not* end the game in round 1. Based on this information, he 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 continuing maximizes expected utility in round 2. Anticipating this, continuing also maximizes expected utility in round 1, as it is likely to get you at least to round 3.

This suggests that the backward induction argument went wrong somewhere. But where? Surely you really ought to end the game in round 99. And surely this means that Bob should end the game in round 98. And so on! This puzzle is sometimes called the **paradox of backward induction**.

Exercise 10.10 ††

Suppose you repeatedly face the Prisoner's 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. What should you do? Does your answer show that you should choose a dominated act?

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 $1/4 \cdot 5 + 3/4 \cdot 0 = 1.25$; for hunting rabbit the utility is of course 1. Individuals going for stag 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. 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. 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. It will lead to 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 an idealisation. 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 picture.

Imagine a population in which individuals repeatedly play a Prisoner's Dilemma wherein they can either cooperate (remain silent, in the original scenario) or defect (confess). Since defectors do better than cooperators in any encounter, it may seem that cooperation can never evolve. On the other hand, 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 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 10.11 †

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

Essay Question 10.1

Explain the paradox of backward induction. Why is it a paradox? How do you think it could be resolved?

Sources and Further Reading

There are many decent introductions to Game Theory. The "Game Theory" entry in the Stanford Encyclopedia by Don Ross (2019) provides a fairly comprehensive overview. A suitable next step might be Steven Tadelis, *Game Theory: An Introduction* (2013).

The paradox of backward induction is discussed, for example, 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). For even more, see Brian Skyrms, "The Stag Hunt and the Evolution of Social Structure" (2004).

11 Bounded Rationality

11.1 Models and reality

We have studied an abstract model of rational agents. The model assumes that an agent has some idea of what the world might be like, which we represent by a credence function Cr over a suitable space of propositions. The agent also has some goals or values or desires, represented by a (possibly partial) utility function U on the same space of propositions. The credence function is assumed to satisfy the formal rules of the probability calculus. It evolves over time by conditionalizing on sensory information, and it satisfies some further constraints like the Probability Coordination Principle. An agent's utility function is assumed to satisfy Jeffrey's axiom, so that it is jointly determined by the agent's credences and their "intrinsic utility function" that assigns a value to the agent's "concerns" – combinations of things the agent ultimately cares about. These intrinsic utilities may in turn be determined by aggregating subvalues. When the agent faces a choice, they are assumed to choose an act that maximizes the credence-weighted average of the utility of the possible outcomes.

Our model is really a family of models, as there are different ways of filling in the details. Should expected utility be understood causally or evidentially? Should credences satisfy some version of the Indifference Principle? Should we rule out some basic desires as irrational? Should we require time consistency? Should we impose constraints on how basic desires may change over time? 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 to which real agents should perhaps aspire. 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 model is descriptively inadequate (no matter how the details are spelled out): that real people

are not expected utility maximizers. In itself, this is not necessarily a problem – not even for the descriptive interpretation of our models. Remember that "all models are wrong". With the possible exception of the standard model of particle physics, the purpose of a model is to identify interesting and robust patterns in the phenomena, not to get every detail right. Nonetheless, it is worth looking at how our model aligns with reality, and what we could change to make it more realistic.

Many supposed cases where people are said to violate the MEU Principle are not counterexamples to the descriptive adequacy of the model we have been studying. Our model can easily accommodate agents who care about risk or fairness or regret (chapter 8). We can accommodate altruistic behaviour (section 1.2), the endowment effect (section 5.2), and apparent failures of time consistency (section 7.4).

Other phenomena are harder to accommodate. People often make mistakes when evaluating the impact of inconclusive information. They don't take into account the "base rate" (section 4.2) or the fact that the information comes from a biased source. They ignore evidence that goes against their opinions.

More simply, most people are bad at maths. Suppose I offer you £100 for telling me the prime factors of 82,717. You have 10 seconds. All you'd have to do, to get the money, is utter '181 and 457'. Moreover, that this is the correct answer logically follows from simpler facts of which you are highly confident. By the rules of probability, you should be confident that '181 and 457' is the correct answer. But you are not.

Exercise 11.1 †††

Explain why, if some proposition C is entailed by two propositions A and B whose probability is greater than 0.99, then the probability of C is greater than 0.98.

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 either agree to a draw or one of them would resign immediately, before the first move. Real people don't play like this.

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 dinner last Monday. As an ideal Bayesian agent, I would still know what I had for dinner on every day of my life.

Exercise 11.2 ††

Show that if an agent conditionalizes on some information E then their credence in E will remain at 1 as long as the agent only changes their beliefs by further applications of conditionalisation. (Conditionalization was introduced in section 4.2.)

There is also indirect evidence that our model does not fit real agents in every respect. The evidence comes from research on artificial intelligence, where our model forms the background for much recent research. Various parts of the model – including the MEU Principle and the Principle of Conditionalization – turn 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 the shop to buy soap. At any point on your way, you could change course. You could decide to turn around, or start running. You could check if your shoe laces are tied. You could mentally compute 181 + 457, or start humming the national anthem. 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 father had decided to run rather than walk to buy soap.) Some authors take the MEU Principle to imply that at each point on your walk,

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 their 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 their 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 this 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 this raises a puzzle. An agent who conforms to our model always chooses acts with greatest expected utility. How are they supposed to do this without calculating? It doesn't seem rational to choose one's acts randomly and maximize expected utility by sheer luck.

Part of the answer is that our model abstracts away from cognitive limitations. Agents who conform to our model have no need to calculate anything. If their evidence entails that a certain act maximizes expected utility, then they are already certain that the act maximizes expected utility: anything that is entailed by their evidence automatically has credence 1.

The idea that expected utility maximizers would constantly have to go through intricate computations also assumes that credences and utilities are conceptually prior to choices. On a preference-based approach, preferences and choices come first. The MEU Principle boils down to certain constraints on preferences, which in turn boil down to constraints on choices. One might hope that even real people, who aren't logically omniscient, can reliably satisfy these constraints without computing expected utilities.

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 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.

In many decision situations, there is no need for sophisticated calculations because one of the acts clearly dominates the others. Whether this is the case depends on the agent's utility function. This suggests that wemight reduce the computational costs of decision-making by tweaking our 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 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 your decision was an intention to go to the shop. Once an intention or plan is formed, we are motivated to execute it. 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've made a mistake when you originally formed the intention.)

Habits can play a similar role. Most of us spend little effort deciding whether we should 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 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 (A and B) 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, $\mathrm{Cr}(\text{Red}) = 1/3$. What would your credences in Green and Blue have to be so that $\mathrm{EU}(A) > \mathrm{EU}(B)$? What would they have to be so that $\mathrm{EU}(D) > \mathrm{EU}(C)$?

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. 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 this matter? One explanation is that people simply prefer lotteries, in which the outcomes have known objective probabilities, to gambles in which the outcomes can only be assigned subjective probabilities. With such a utility function,

the outcome labelled '£1000' in A is actually better than the corresponding outcome in C, because only the former involves having chosen a lottery.

But why would agents prefer lotteries? A possible answer is that such a preference tends to reduce computational costs. If you know the objective probabilities of a state, it is easy to figure out the credence you should give to the state: it should match the objective probabilities. If you don't know the objective probability, more work may be required to figure out the extent to which the state is supported by your total evidence. 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 (as is the expected payoff of C and D); a preference for lotteries tips the balance in favour of 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 want to design a robot – an artificial agent with a probabilistic representation of its environment and some goals. Let's assume that we want our agent to assign credences and utilities to a total of 50 logically independent propositions $A_1, \dots, A_{50}$ (an absurdly small number). How large of a database do we need?

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 $\mathrm{U}(A \wedge B)$ or $\mathrm{U}(A \vee B)$ from $\mathrm{U}(A)$ and $\mathrm{U}(B)$. If we want to determine the agent's entire credence and utility functions (without further assumptions), we need to store at least the probability and utility of every "possible world" – every maximally consistent conjunction of $A_1, \dots, A_{50}$ and their negations.

Exercise 11.6 †††

Explain why the probability of every proposition that can be defined in terms of $A_1, \dots, A_{50}$ can be computed from the probability assigned to these "worlds". Then explain why the utility of every such proposition can be com-

puted 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 both credences and utilities, we need a database with $2,251,799,813,685,248$ records. (I am exaggerating. 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 robot 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 chapters 5.4 and 7, we have encountered two tricks that allow us to simplify the representation of an agent's utility function. First, if the agent cares about some attributes 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 attributes they care about (section 5.4). If, for example, our robot 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, if our robot's preferences are separable with respect to these attributes, then the value of any combination of the 20 propositions and their negations can be determined by adding up relevant subvalues (section 7.2). We can cut down the number of utility records from 2^{20} to $2 \cdot 20 = 40$.

Similar tricks are available for the agent's credence function. Mirroring the first trick, we could explicitly store only the robot's credence in certain sets of worlds, and assume that its credence is distributed uniformly within these sets. The trick can be extended to non-uniform distributions. For example, suppose our robot has imperfect information about how far it is from the next charging station. Instead of explicitly storing a probability for every possible distance (1m, 2m, 3m, …), we might assume that the robot's credence over these possibilities follows a Gaussian 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 sometimes plausible. Whether the next charging station is 100 meters away plausibly doesn't depend on whether the outside temperature is above 20°C. For many other propositions, however, independence is implausible. On the supposition that it is warm outside (W), it may well be more likely that the window is open (O), or that there are people on the street (P), than on the supposition that it isn't warm ($\neg W$). If our agent is unsure whether it is warm, it follows that $\mathrm{Cr}(O/W) > \mathrm{Cr}(O)$, and $\mathrm{Cr}(P/W) > \mathrm{Cr}(P)$. 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**. 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: $\mathrm{Cr}(P/O) > \mathrm{Cr}(P)$. So P and O 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:

$$\mathrm{Cr}(P/O \wedge W) = \mathrm{Cr}(P/W).$$

In this case, we say that P and O are independent *conditional on* W.

Now consider the possible combinations of W, P, O and their negations. By the probability calculus (compare exercise 2.10),

$$\mathrm{Cr}(W \wedge P \wedge O) = \mathrm{Cr}(W) \cdot \mathrm{Cr}(O/W) \cdot \mathrm{Cr}(P/O \wedge W).$$

By the above assumption of conditional independence, this simplifies to

$$\mathrm{Cr}(W \wedge P \wedge O) = \mathrm{Cr}(W) \cdot \mathrm{Cr}(O/W) \cdot \mathrm{Cr}(P/W).$$

In general, with the assumption of conditional independence, we can fix the probability of all combinations of W, P, O, and their negations by specifying the probability of W, the probability of P conditional on W and on $\neg W$, and the probability of O conditional on W and on $\neg W$. The number of required records shrinks from $2^3 - 1 = 7$ to 5. This may not look all that impressive, but the method really pays off if more than three propositions are involved.

The present technique for exploiting conditional independence to simplify proba-

bilistic models is formalized in the theory 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 widely used in artificial intelligence to model decision-making agents.

A decision maker needs information not only about the present state of the world, but also about the future. We can represent a history of states as a sequence S_1, S_2, S_3,..., 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 each of these possibilities, 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 usually assume that the decision maker's intrinsic preferences are stationary and separable across times, so that the value of a history of states is a discounted sum of a 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.

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. It often helps to ignore states in the distant future and let the agent maximize the expected total utility in the next few states only. Several techniques have been developed that allow an efficient *approximate* computation of expected utilities and posterior probabilities. These techniques are often supplemented by a meta-decision process that 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 have been argued to be a consequence 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 in the 1970s-1990s.

Prospect theory has to be understood on the background of a highly restricted version of decision theory that dominates economics. The highly restricted theory assumes that utility is only defined for money and other material goods, and it only deals with choices between lotteries, where the objective probabilities are known. People are assumed to want more money and goods, but 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. 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 0.95. 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 an alternative to the more liberal model that we have been studying. 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 people don't actually do this. 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. From this perspective, prospect theory assumes that people systematically violate the Probability Coordination Principle. Their credence in low-probability events 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. But there is evidence that people really do fail to coordinate their beliefs with known objective probabilities, especially if the probabilities are communicated verbally. People's decision weights tend to be closer to the objective probabilities if they have experienced the probabilities as relative frequencies in repeated trials.

Reference dependence may also raise a genuine challenge. Many forms of reference dependence can easily be accommodated in our model. We can allow that

people care about how much they own in comparison to what they have owned before, or in comparison to what their peers own. But sometimes the reference point is affected by intuitively irrelevant features of the context, and this is harder to square with our model.

Exercise 11.7 †

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 related to so-called **framing effects**. In experiments, people's choices can systematically depend on how one and the same decision problem is described. 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 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. On reflection, however, 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. This 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 take this kind of attribute selection

into account.

11.5 Imprecise credence and utility

One respect in which our model is often found unrealistic is that its credences and utilities are too precise. What is your credence that there will be a nuclear war before 2100? Is it 0.31832? Or 0.20993? Any such answer may seem wrong. You haven't made up your mind up to the 5th (let alone the 500th) decimal point.

Across several disciplines, researchers have developed models that don't assume unique and precise credences and utilities. On this approach, your credence in a nuclear war might be given by a whole range of numbers – perhaps by the interval [0.2, 0.5] that contains all numbers from 0.2 up to 0.5.

If we want to specify rational norms for such "imprecise" credences, it helps to assume that they are determined by a set of precise credence functions. We would model your imprecise belief state by a set $\mathbb{Cr}$ of credence functions that assign different numbers to the nuclear war hypothesis. For each number in [0.2, 0.5], there would be a member of $\mathbb{Cr}$ that assigns this number to the nuclear war hypothesis. We can then implicitly constrain your imprecise credences by saying that each member of $\mathbb{Cr}$ should satisfy the Kolmogorov axioms.

We can adopt a similar account of utility, replacing our single utility function U by a set of utility functions $\mathbb{U}$.

On a preference-based approach, "imprecise" credences and utilities naturally arise through violations of the Completeness axiom. Completeness says that for any propositions A and B, you either prefer A to B, or you prefer B to A, or you are indifferent between A and B. This is trivial if we define preference in terms of choice. Indeed, in a forced choice between A and B, you will inevitably choose either A or B; even indifference can be ruled out. But we've 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. Once we take a step back from choice behaviour, it is conceivable that you neither prefer A to B, nor B to A, and yet you're not indifferent between the two. You simply haven't 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 this mean that you're exactly indifferent? Not necessarily. If you were, you should definitely prefer losing the capacity to hear *and getting £10* to losing the capacity to walk. In reality, the added £10 may not make a difference.

Exercise 11.8 ††

Suppose we define '$A \sim B$' as 'not $A \succ B$ and not $B \succ A$'. It is then logically guaranteed that either $A \succ B$ or $B \succ A$ or $A \sim B$. But Transitivity might fail, if you haven't fully made up your mind. Explain why.

Even if we give up completeness, however, we might still require **completability**. We might want to say that if an agent's preferences violate, say, Ramsey's axioms because they fail to rank certain options, then there is a refinement of their preferences, filling in the missing rankings, that does satisfy the axioms. Ramsey's representation theorem then implies that the agent's preferences are represented by a *set* of credence and utility functions.

Allowing for a set of credence and utility functions requires some changes to our model. How should a set of credence functions be revised when new information comes in? How should an agent choose based on a set of credence and utility functions? Both questions raise serious problems.

The most obvious answer to the first question is that if an agent has a set of credence functions $\mathbb{C}\mathrm{r}$ and receives total evidence E, then her new set of credence functions should result from $\mathbb{C}\mathrm{r}$ by conditionalising each member of $\mathbb{C}\mathrm{r}$ on E.

One problem with this answer is that this process is, in general, computationally *harder* than conditionalising a single probability measure. In this respect, our model has become less realistic, not more.

Here is another problem. Suppose I have an urn containing 2 balls, one of which is white. The other is either white or red. You have no opinion about how the other ball's colour: your belief state $\mathbb{C}\mathrm{r}$ contains all possible probability assignments to the hypothesis that the other ball is white. Now I shuffle the urn, draw a ball, and show it to you. The ball is white. If you conditionalise each member of $\mathbb{C}\mathrm{r}$ on this information, your belief state remains unchanged! Your new imprecise credence is still $\mathbb{C}\mathrm{r}$. It remains at $\mathbb{C}\mathrm{r}$ no matter how often I draw a white ball, each time replacing the previously drawn ball. This seems wrong.

Exercise 11.9 †††

Explain why seeing a white ball doesn't change $\mathbb{C}\mathrm{r}$.

Let's briefly turn to the other question. How should you choose between some options if you have a set of credence and utility functions? Suppose option A maximizes expected utility relative to one of your credence and utility functions, while option B maximizes expected utility according to another. Should you choose A or B? A popular "permissivist" answer is that either choice is acceptable.

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.

But now imagine you are offered two bets A and B, one after the other, on a proposition H to which you don't assign a precise credence. Let's say your credence in H spans the range from 0.2 to 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 $\neg H$. Assume for simplicity that your utility is precise and proportional to the monetary payoff. The permissivst account then classifies both bets as optional: you may take them or leave them. But accepting both bets yields a guaranteed gain of £0.40. By refusing both bets, you would miss out on a sure gain.

Sources and Further Reading

A standard textbook on artificial intelligence is Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (4th ed., 2020). Part IV covers most of the material I have summarized in section 11.3.

For evidence that our brains might use some of the tricks AI researchers have found see, for example, Samuel Gershman and Nathaniel Daw, "Perception, Action and Utility" (2012). For a more high-level view on the idea that cognitive systems try to approximate the Bayesian ideal, see Thomas Griffiths et al, "Rational Use of Cognitive Resources" (2015), or Samuel Gershman et al, "Computational rationality: A

converging paradigm" (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 empirical claims about probabilities, frequencies, and reference points in section 11.4 are from Kahneman's *Thinking Fast and Slow* (2011).

For a model of attribute selection in the evaluation of options, see Franz Dietrich and Christian List, "Reason-Based Choice and Context-Dependence" (2016). There are also models for how to selectively use different aspects of a credence function. See Peter Fritz and Harvey Lederman, "Standard State Space Models of Unawareness" (2015).

The Stanford Encyclopedia Entry "Imprecise Probabilities" by Saemus Bradley (2019) provides a good overview of research on the topic of section 11.5. The urn problem is an instance of "belief inertia".

The Ellsberg Paradox was presented in Daniel Ellsberg, "Risk, Ambiguity, and the Savage Axioms" (1961).