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 Mushroon 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 Mushroon 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 Mushroon 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'.

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).