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. Suppose I offer you £100 for telling me the prime factors of 82,717. You have 10 seconds. Can you do it? Probably not. All you'd have to do, to get the money, is utter '181 and 457', which is surely an available act. Moreover, that '181 and 457' is the correct answer logically follows from simpler facts of which you are highly confident. By the rules of probability, you should be confident that '181 and 457' is the correct answer. As an expected utility maximizer (assuming you'd like to get the £100), you would utter these words. Yet you don't.

Exercise 11.1 ††

Explain why, if some proposition C logically follows from two other propositions A and B, and $\mathrm{Cr}(A) > 0.9$ and $\mathrm{Cr}(B) > 0.9$, then $\mathrm{Cr}(C) > 0.81$.

In 1913, Ernst Zermelo proved that in the game of chess, there is either a strategy for the starting player, White, that guarantees victory no matter what Black does, or there is such a strategy for Black, or there is a strategy for either player to force a draw. Consequently, if two ideal Bayesian agents sat down to a game of chess, and their only interest was in winning, they would 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 $\text{Cr}_t(A) = 1$, and the agent conditionalizes on information E with $\text{Cr}_t(E) > 0$, then $\text{Cr}_{t+1}(A) = 1$. (Conditionalization was introduced in section 4.2.)

There is also strong 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 mother 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 in many circumstances, simple alternatives to computing expected utilities reliably lead to optimal choices. As the psychologist Gerd Gigerenzer once pointed out, if you want to catch a flying ball, an efficient alternative to computing the ball's trajectory – which is generally intractable – is to move around in such a way that the angle between you and the ball remains within a certain range. This ensures that you'll eventually stand where the ball will arrive.

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.

One reason why many decision problems don't require sophisticated computations is that one of the acts clearly dominates the others. Whether this is the case depends on the agent's utility function. It follows that you can reduce the computational costs of decision-making by tweaking your utilities.

For example, suppose you assign significant (sub)value to obeying orders. Doing whatever you're ordered to do is then a reliable way of maximizing expected utility, and it requires 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, $\mathrm{Cr}(\mathit{Red}) = 1/3$. What would your credences in *Green* and *Blue* have to look like so that $\mathrm{EU}(A) > \mathrm{EU}(B)$? What would they have to look like 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 $Cr(A \land B)$ or $\mathrm{Cr}(A \lor B)$ from $\mathrm{Cr}(A)$ and $\mathrm{Cr}(B)$. Nor can we compute $\mathrm{U}(A \land B)$ or $\mathrm{U}(A \lor B)$ 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 computed from the probability and utility assigned to the worlds.

There are $2^{50} = 1,125,899,906,842,624$ maximally consistent conjunctions of $A_1, \dots, A_{50}$ 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 $A_1, \dots, A_n$, 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 (1 m, 2 m, 3 m, …), 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 $\text{Cr}(A \land B) = \text{Cr}(A) \cdot \text{Cr}(B)$. If all the 50 propositions $A_1, \dots, A_{50}$ 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, it may well be more likely that the window is open, or that there are people on the street, than on the supposition that it isn't warm. If our agent is unsure whether it is warm, it follows that $\text{Cr}(\textit{Open}/\textit{Warm}) > \text{Cr}(\textit{Open})$, and $\text{Cr}(\textit{People}/\textit{Warm}) > \text{Cr}(\textit{People})$. We can't assume probabilistic independence across all the 50 propositions $A_1, \dots, A_{50}$.

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: $\mathrm{Cr}(\mathit{People}/\mathit{Open}) > \mathrm{Cr}(\mathit{People})$. So *People* and *Open* are not independent. However, *on the supposition that it is warm outside*, the window being open may no longer increase the probability of people on the street:

$$\mathrm{Cr}(\mathit{People}/\mathit{Open} \land \mathit{Warm}) = \mathrm{Cr}(\mathit{People}/\mathit{Warm}).$$

In this case, we say that *People* and *Open* are independent *conditional on Warm*.

Now consider the possible combinations of *Warm*, *People*, *Open* and their negations. By the probability calculus (compare exercise 2.10),

$$\mathrm{Cr}(\mathit{Warm}\land\mathit{People}\land\mathit{Open}) = \mathrm{Cr}(\mathit{Warm})\cdot\mathrm{Cr}(\mathit{Open}/\mathit{Warm})\cdot\mathrm{Cr}(\mathit{People}/\mathit{Open}\land\mathit{Warm}).$$

By the above assumption of conditional independence, this simplifies to

$$\mathrm{Cr}(\mathit{Warm} \land \mathit{People} \land \mathit{Open}) = \mathrm{Cr}(\mathit{Warm}) \cdot \mathrm{Cr}(\mathit{Open}/\mathit{Warm}) \cdot \mathrm{Cr}(\mathit{People}/\mathit{Warm}).$$

In general, with the assumption of conditional independence, we can fix the probability of all combinations of *Warm*, *People*, *Open*, and their negations by specifying the probability of *Warm*, the probability of *People* conditional on *Warm* and on $\neg$*Warm*, and the probability of *Open* conditional on *Warm* and on $\neg$*Warm*. 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 probabilistic 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 $\mathrm{Cr}(S_3/S_1 \land S_2) = \mathrm{Cr}(S_3/S_2)$. 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 0.92. Among other things, this is meant to explain why people play the lottery, and why they tend to pay a high price for certainty: they prefer a settlement of £90000 over a trial in which they have a 99% chance of getting £100000 but a 1% chance of getting nothing.

Prospect theory is clearly an alternative to the simplistic economical model mentioned above. It is not so obvious whether it is 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 closely related to so-called **framing effects**. In experiments, people's choices can systematically depend on how one and the same decision problem is described. 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

I'm going to toss three dice. Would you rather get £1000 if the sum of the numbers on the dice is at least 10 or if all three numbers are different? You'd probably need some time to give a final answer. You know that the six possible results for every dice have equal probability, and that the results are independent. But it takes some effort to figure out which of the two events I described is more likely.

A real agent's cognitive system can't explicitly store a credence and utility for every proposition. It can only store a limited number of **constraints** on credences and utilities. A constraint rules out some credences and utilities, but not others. That outcomes of die tosses are probabilistically independent is a constraint; it entails that the probability of three sixes is the product of the probabilities for the individual dice: $\mathrm{Cr}(\mathrm{Six}_1 \land \mathrm{Six}_2 \land \mathrm{Six}_3) = \mathrm{Cr}(\mathrm{Six}_1) \cdot \mathrm{Cr}(\mathrm{Six}_2) \cdot \mathrm{Cr}(\mathrm{Six}_3)$, but it does not fix what these probabilities are.

Will the constraints stored by an agent's cognitive system always be rich enough to determine a unique credence and utility function? Perhaps not. There might well be questions on which you don't have a settled opinion, even in principle, after ideal reflection. Or suppose you don't have time for lengthy reflection, or you're too tired. It might be wrong to model your attitudes by a single, precise credence function, and a single, precise utility functions.

Across several disciplines, researchers have developed models that don't assume unique and precise credences and utilities. The standard approach is to use **sets of credence and utility functions** instead of single functions. If we think of your cognitive system as storing constraints, then the set comprises all the (pairs of) functions that meet these constraints. Intuitively, each member of the set is a *refinement* or *precisification* of your indeterminate state of mind.

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 seems perfectly possible 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.

Here is a possible 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 £1* to losing the capacity to walk. In reality, the added £1 may not make a difference.

Exercise 11.8 ††

Suppose we define '$A \sim B$' as 'not $A \succ B$ and not $B \succ A$'. Completeness is then logically guaranteed. 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 $\mathbb{Cr}$ and receives total evidence E, then her new set of credence functions should result from $\mathbb{Cr}$ by conditionalising each member of $\mathbb{Cr}$ 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 $\mathbb{Cr}$ 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 $\mathbb{Cr}$ on this information, your belief state remains unchanged! Your new imprecise credence is still $\mathbb{Cr}$. It remains at $\mathbb{Cr}$ 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 $\mathbb{Cr}$.

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).