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. Two propositions A and B are **logically independent** if all the combinations $A \wedge B$, $A \wedge \neg B$, $\neg A \wedge B$, and $\neg A \wedge \neg B$ are logically possible. A and B are **probabilistically independent** relative to some credence function Cr if $\mathrm{Cr}(B/A) = \mathrm{Cr}(B)$. (See section 2.4.) A and B are **causally independent** if, whether one of them is true has no causal influence on whether the other 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 highlight some implications of the theory from section 5.3. Let $S_1, \ldots, S_n$ be a collection of mutually incompatible and jointly exhaustive propositions, so that exactly one of them is true at every possible world. Such a collection is called a **partition**. Jeffrey's axiom entails that for any proposition A with $\mathrm{Cr}(A) > 0$,

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

If we think of propositions as regions in logical space, then each $A \wedge S_i$ is a subregion of A. (J1) says that the desirability of the whole region is a weighted average of the desirability of its parts, weighted by their probability conditional on A. This could be contested, but even causal decision theorists tend to agree.

Equation (J1) holds for every partition $S_1, \ldots, S_n$. Now consider partitions $S_1, \ldots, S_n$ that are fine-grained enough so that every conjunction of A with a member of the partition settles everything an 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

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

It is now easy to see why an act's evidential expected utility equals its utility. 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 these states. The evidential expected utility of an act A is

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

Each conjunction of an act A with one of the states S_i settles everything the agent cares about. So equation (J2) applies. Also, 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 $\mathrm{EU_e}$.

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 ($U = 1, \mathrm{Cr} = 0.9$)	Fail & No Fun ($U = -8, \mathrm{Cr} = 0.1$)
Don't Study	Pass & Fun ($U = 5, \mathrm{Cr} = 0.2$)	Fail & Fun ($U = -2, \mathrm{Cr} = 0.8$)

'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

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

The unconditional credences $\mathrm{Cr}(S_i)$ in the old method have been replaced by conditional credences $\mathrm{Cr}(S_i/A)$, 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 evidential 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.

Exercise 9.4 †††

Above I assumed that the 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 black box and a transparent box. The transparent box contains £1000. You can't see what's in the black box. You have two options. You can take *just the black box* and keep whatever is inside. Alternatively, you can take *both boxes* and keep their content. A demon has tried to predict what you will do. If she has predicted that you will take both boxes, then she has put nothing in the black box. If she has predicted that you will take just the black 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 black box	£0 in black box
Take only black box	£1,000,000	£0
Take both boxes	£1,001,000	£1000

Note that the states are causally independent of the acts, as CDT requires. Whether you take both boxes or just the black box – in philosophy jargon, whether you *two-box* or *one-box* – is certain to have no causal influence over what's in the boxes. This is crucial to understanding Newcomb's Problem. By the time of your choice, the content of the boxes is settled. The demon won't magically change what's in the black box in response to your choice. Her only superpower is predicting people's choices.

It is obvious from the decision matrix that taking both boxes maximizes causal expected utility, since it dominates one-boxing: it is better in every state. We don't need to fill in the precise utilities and probabilities.

Turning to EDT, we do need to specify a few more details. Let's say you are 95% confident that there is a million in the black box if you one-box, and 5% confident that there is a million in the black box if you two-box. Let's also assume (for simplicity) that your utility is proportional to the amount of money you will get. Using the "new method" 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):

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

One-boxing comes out better than two-boxing.

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

Some think one-boxing is obviously right. You're almost certain to get more if you one-box than if you two-box. Look at all the people that have been offered the choice in the past! Those who one-boxed almost always walked away with a million, while those who two-boxed walked away with a thousand. Wouldn't you rather be in the first group than in the second? It's your choice!

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 almost certain to get you a million, two-boxing 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 black box. Remember that the content of the boxes is settled. The black box either contains a thousand or a million. One-boxing and two-boxing both give you the black box. It is settled that you will get however much is in that box. The only thing that isn't settled – the only thing over which you have any control – is whether you also get the £1000 from the transparent box. And if you prefer more money to less money, then clearly (so the argument) you should take the additional £1000.

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 black box. You've agreed to a secret signal by which she will let you know whether it would be better for you to choose both boxes or just the black box. If you trust your friend, it seems that you should follow her advice. But what will she signal? If the box is empty, she will 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 black 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 ††

Show that if you follow EDT then you would not want to know what's in the black box. You'd be willing to pay the demon £500 for not revealing to you the content of the box.

What about the fact that one-boxers are generally richer than two-boxers? Doesn't this show that the one-boxers are doing something right? Not so, say those who advocate two-boxing. The two-boxers who walked away with a mere thousand were never given a chance to get a million. They were confronted with an empty black box and a transparent box containing £1000; it's hardly their fault that they didn't get a million. All those one-boxers who got a million were effectively given a choice between £1,001,000 and £1,000,000. The fact that they got a million hardly shows that they made the right choice. As an analogy, imagine there are two buttons labelled 'dark' and 'blonde'. If you press the button that matches your hair colour, you get a million if your hair is blonde and a thousand if it is dark. Almost everyone who presses 'blonde' walks away with a million, while almost everyone who presses 'dark' walks away with a thousand. It doesn't follow that everyone should have pressed 'blonde'. Those with dark hair never had a chance to get a million.

9.3 More realistic Newcomb Problems?

Newcomb's Problem is science fiction. Nobody ever faces that situation. Why should we care about the answer?

Philosophers care because the problem brings to light a more general issue: whether

the norms of practical rationality involve causal notions. Those who favour two-boxing in Newcomb's Problem argue that the apparent advantage of EDT, that it does not appeal to causal notions, is actually a flaw.

In effect, EDT recommends choosing acts whose choice would be good news. One-boxing in Newcomb's Problem would be good news because it would provide strong evidence that the black box (which you're certain to get) contains a million. 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 black 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 black 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 (sub)value 100 to A winning and 0 to B winning. Voting would add a (sub)value of -1, since it would cause you some inconvenience. Your credence that your vote will make a difference is 0.001. You figure out that not voting maximizes expected utility. But then you realize that other potential voters are likely to go through the same thought process as you. You estimate that around 1% of A's supporters might go through the same process of deliberation as you, and will reach the same conclusion that you will reach. Does this change the causal expected utility of voting? Does it change the evidential expected utility? (Explain briefly, without computing anything.)

9.4 Causal decision theories

Those who are convinced by the case against EDT believe that some causal notion must figure in an adequate theory of rational choice: rational agents maximize causal expected utility.

One way to define causal 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 decide 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?

Exercise 9.11 ††

What should you do in the scenario from exercise 9.2, assuming CDT?

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 black box is noted in Brian Skyrms, "Causal Decision Theory" (1982). The hair colour button is from Adam Bales, *Decision and Dependence* (2017). 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).