Chapter 18 Causal Decision Theory

18.1 Causal and Evidential Decision Theory

Over the last two chapters we've looked at two ways of thinking about the expected utility of an action A. These are

$$Pr(S_1)U(S_1A) + ... + Pr(S_n)U(S_nA)$$
$$Pr(S_1|A)U(S_1A) + ... + Pr(S_n|A)U(S_nA)$$

It will be convenient to have names for these two approaches. So let's say that the first of these, which uses unconditional probabilities, is **causal expected value**, and the second of these, which uses conditional probabilities is the **evidential expected value**. The reason for the names should be clear enough. The causal expected value measures what you can expect to bring about by your action. The evidential expected value measures what kind of result your action is evidence that you'll get.

Causal Decision Theory then is the theory that rational agents aim to maximise causal expected utility.

Evidential Decision Theory is the theory that rational agents aim to maximise evidential expected utility.

Over the past two chapters we've been looking at reasons why we should be causal decision theorists rather than evidential decision theorists. We'll close out this section by looking at various puzzles for causal decision theory, and then looking at one reason why we might want some kind of hybrid approach.

18.2 Right and Wrong Tabulations

If we use the causal approach, it is very important how we divide up the states. We can see this by thinking again about an example from Jim Joyce that we discussed a while ago.

> Suupose you have just parked in a seedy neighborhood when a man approaches and offers to "protect" your car from harm for \$10. You recognize this as extortion and have heard that people who refuse "protection" invariably return to find their windshields smashed. Those who pay find their cars intact. You cannot park anywhere else because you are late for an important meeting. It costs \$400 to replace a windshield. Should you buy "protection"? Dominance says that you should not. Since you would rather have the extra \$10 both in the even that your windshield is smashed and in the event that it is not, Dominance tells you not to pay. (Joyce, *The Foundations of Causal Decision Theory*, pp 115-6.)

If we set this up as a table, we get the following possible states and outcomes.

	Broken Windshield	Unbroken Windshield
Pay extortion	-\$410	-\$10
Don't pay	-\$400	0

Now if you look at the causal expected value of each action, the expected value of not paying will be higher. And this will be so whatever probabilities you assign to broken windshield and unbroken windshield. Say that the probability of the first is x and of the second is $1-x$. Then we'll have the following (assuming dollars equal utils)

$$\begin{aligned}
Exp(U(\text{Pay extortion})) &= -410x - 10(1-x) \\
&= -400x - 10 \\
Exp(U(\text{Don't pay}) &= -400x - 0(1-x) \\
&= -400x
\end{aligned}$$

Whatever x is, the causal expected value of not paying is higher by 10. That's obviously a bad result. Is it a problem for causal decision theory though? No. As the name 'causal' suggests, it is crucial to causal decision theory that we separate out what we have causal power over from what we don't have causal power over. The states of the world represent what we can't control. If something can be causally affected by our actions, it can't be a background state.

So this is a complication in applying causal decision theory. Note that it is not a problem for evidential decision theory. We can even use the very table that we have there. Let's assume that the probability of broken windshield given paying is 0, and the probability of unbroken windshield given paying is 0. Then the expected utilities will work out as follows

$$\begin{aligned}
Exp(U(\text{Pay extortion})) &= -410 \times 0 - 10 \times 1 \\
&= -10 \\
Exp(U(\text{Don't pay}) &= -400 \times 1 - 10 \times 0 \\
&= -400
\end{aligned}$$

So we get the right result that we should pay up. It is a nice feature of evidential decision theory that we don't have to be so careful about what states are and aren't under our control. Of course, if the only reason we don't have to worry about what is and isn't under our control is that the theory systematically ignores such facts, even though they are intuitively relevant to decision theory, this isn't perhaps the best advertisement for evidential decision theory.

18.3 Why Ain'Cha Rich

There is one other argument for evidential decision theory that we haven't yet addressed. Causal decision theory recommends taking two boxes in Newcomb's problem; evidential decision theory recommends only taking one. People who take both boxes tend, as a rule, to end up poorer than people who take just the one box. Since the aim here is to get the best outcome, this might be thought to be embarrassing for causal decision theorists.

Causal decision theorists have a response to this argument. They say that Newcomb's problem is a situation where there is someone who is quite smart, and quite determined to reward irrationality. In such a case, they say, it isn't too surprising that irrational people, i.e. evidential decision theorists, get rewarded. Moreover, if a rational person like them were to have taken just one box, they would have ended up with even less money, i.e., they would have ended up with nothing.

One way that causal decision theorists would have liked to make this objection stronger would be to show that there is a universal problem for decision theories - whenever there is someone whose aim is to reward people who don't follow the dictates of their theory, then the followers of their theory will end up poorer than the non-followers. That's what happens to causal decision theorists in Newcomb's problem. It turns out it is hard, however, to play such a trick on evidential decision theorists.

Of course we could have someone go around and just give money to people who have done irrational things. That wouldn't be any sign that the theory is wrong however. What's distinctive about Newcomb's problem is that we know this person is out there, rewarding non-followers of causal decision theory, and yet the causal decision theorist does not change their recommendation. In this respect they differ from evidential decision theorists.

It turns out to be very hard, perhaps impossible, to construct a problem of this sort for evidential decision theorists. That is, it turns out to be hard to construct a problem where (a) an agent aims to enrich all and only those who don't follow evidential decision theory, (b) other agents know what the devious agent is doing, but (c) evidential decision theory still ends up recommending that you side with those who end up getting less money. If the devious agent rewards doing X, then evidential decision theory will (other things equal) recommend doing X. The devious agent will make such a large evidential difference that evidential decision theory will recommend doing the thing the devious agent is rewarding.

So there's no simple response to the "Why Ain'Cha Rich" rhetorical question. The causal decision theorist says it is because there is a devious agent rewarding irrationality. The evidential decision theorist says that a theory should not allow the existence of such an agent. This seems to be a standoff.

18.4 Dilemmas

Consider the following story, told by Allan Gibbard and William Harper in their paper setting out causal decision theory.

> Consider the story of the man who met Death in Damascus. Death looked surprised, but then recovered his ghastly composure and said, 'I AM COMING FOR YOU TOMORROW'. The terrified man that night bought a camel and rode to Aleppo. The next day, Death knocked on the door of the room where he was hiding, and said 'I HAVE COME FOR YOU'.
>
> 'But I thought you would be looking for me in Damascus', said the man.
>
> 'NOT AT ALL', said Death 'THAT IS WHY I WAS SURPRISED TO SEE YOU YESTERDAY. I KNEW THAT TODAY I WAS TO FIND YOU IN ALEPPO'.
>
> Now suppose the man knows the following. Death works from an appointment book which states time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with Death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo...
>
> If... he decides to go to Aleppo, he then has strong grounds for expecting that Aleppo is where Death already expects him to be, and hence it is rational for him to prefer staying in Damascus. Similarly, deciding to stay in Damascus would give him strong grounds for thinking that he ought to go to Aleppo.

In cases like this, the agent is in a real dilemma. Whatever he does, it seems that it will be the wrong thing. If he goes to Aleppo, then Death will probably be there. And if he stays in Damascus, then Death will probably be there as well. So it seems like he is stuck.

Of course in one sense, there is clearly a right thing to do, namely go wherever Death isn't. But that isn't the sense of right decision we're typically using in decision theory. Is there something that he can do that maximises expected utility. In a sense the answer is "No". Whatever he does, doing that will be some evidence that Death is elsewhere. And what he should do is go wherever his evidence suggests Death isn't. This turns out to be impossible, so the agent is bound not to do the rational thing.

Is this a problem for causal decision theory? It is if you think that we should always have a rational option available to us. If you think that 'rational' here is a kind of 'ought', and you think 'ought' implies 'can', then you might think we have a problem, because in this case there's a sense in which the man can't do the right thing. (Though this is a bit unclear; in the actual story, there's a perfectly good sense in which he could have stayed in Aleppo, and the right thing to do, given his evidence, would have been to stay in Aleppo. So in one sense he could have done the right thing.) But both the premises of the little argument here are somewhat contentious. It isn't clear that we should say you ought, in any sense, maximise expected utility. And the principle that ought implies can is rather controversial. So perhaps this isn't a clear counterexample to causal decision theory.

18.5 Weak Newcomb Problems

Imagine a small change to the original Newcomb problem. Instead of there being $1000 in the clear box, there is $800,000. Still, evidential decision theory recommends taking one box. The evidential expected value of taking both boxes is now roughly $800,000, while the evidential expected value of taking just the one box is $1,000,000. Causal decision theory recommends taking both boxes, as before.

So neither theory changes its recommendations when we increase the amount in the clear box. But I think many people find the case for taking just the one box to be less compelling in this variant. Does that suggest we need a third theory, other than just causal or evidential decision theory?

It turns out that we can come up with hybrid theories that recommend taking one box in the original case, but two boxes in the original case. Remember that in principle anything can have a probability, including theories of decision. So let's pretend that given the (philosophical) evidence on the table, the probability of causal decision theory is, say, 0.8, while the probability of evidential decision theory is 0.2. (I'm not saying these numbers are right, this is just a possibility to float.) And let's say that we should do the thing that has the highest *expected* expected utility, where we work out expected expected utilities by summing over the expectation of the action on different theories, times the probability of each theory. (Again, I'm not endorsing this, just floating it.)

Now in the original Newcomb problem, evidential decision theory says taking one boxes is $999,000 better, while causal decision theory say staking both boxes is $1,000 better. So the expected expected utility of taking one box rather than both boxes is $0.2 \times 999,000 - 0.8 \times 1,000$, which is 199,000. So taking one box is 'better' by 199,000

In the modified Newcomb problem, evidential decision theory says taking one boxes is $200,000 better, while causal decision theory says taking both boxes is $800,000 better. So the expected expected utility of taking one box rather than both boxes is $0.2 \times 200,000 - 0.8 \times 800,000$, i.e., -600,000. So taking both boxes is 'better' by 600,000.

If you think that changing the amount in the clear box can change your decision in Newcomb's problem, then possibly you want a hybrid theory, perhaps like the one floated here.