Chapter 17 Realistic Newcomb Problems

17.1 Real Life Newcomb Cases

In the previous notes we ended up saying that there are two quite different ways to think about utility expectations. We can use the unconditional probability of each state, or, for each choice, we can use the probabilities of each state conditional on the choice the agent makes. That is, we can take the expected utility of a choice A to be given by one or other of the following formulae.

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

It would be nice to know which of these is the right formula, since the two formulae disagree about cases like Newcomb's problem. Since we have a case where they disagree, a simple methodology suggests itself. Figure out what we should do in Newcomb's problem, and then select the formula which agrees with the correct answer. But this method has two flaws.

First, intuitions about Newcomb's puzzle are themselves all over the place. If we try to adjust our theory to match our judgments in Newcomb's problem, then different people will have different theories.

Second, Newcomb's problem is itself quite fantastic. This is part of why different people have such divergent intuitions on the example. But it also might make us think that the problem is not particularly urgent. If the two equations only come apart in fantastic cases like this, perhaps we can ignore the puzzles.

So it would be useful to come up with more realistic examples where the two equations come apart. It turns out that what is driving the divergence between the equations is that there is a common cause of the world being in a certain state and you making the choice that you make. Any time there is something in the world that tracks your decision making processes, we'll have a Newcomb like problem.

For example, imagine that we are in a Prisoners' Dilemma situation where we know that the other prisoner uses very similar decision making procedures to what we use. Here is the table for a Prisoners' Dilemma.

	Other Cooperates	Other Defects
You Cooperate	(3,3)	(0, 5)
You Defect	(5, 0)	(1, 1)

In this table the notation (x, y) means that you get x utils and the other person gets y utils. Remember that utils are meant to be an overall measure of what you value, so it includes your altruistic care for the other person.

Let's see why this resembles a Newcomb problem. Assume that conditional on your performing an action A, the probability that the other person will do the same action is 0.9. Then, if we are taking probabilities to be conditional

on choices, the expected utility of the two choices is

$$\begin{aligned} Exp(U(Coop)) &= 0.9 \times 3 + 0.1 \times 0 \\ &= 2.7 \\ Exp(U(Defect)) &= 0.1 \times 5 + 0.9 \times 1 \\ &= 1.4 \end{aligned}$$

So if we use probabilities conditional on choices, we end up with the result that you should cooperate. But note that cooperation is dominated by defection. If the other person defects, then your choice is to get 1 (by defecting) or 0 (by cooperating). You're better off cooperating. If the other person cooperates, then your choice is to get 5 (by defecting) or 0 (by cooperating). So whatever probability we give to the possible actions of the other person, provided we don't conditionalise on our choice, we'll end up deciding to defect.

Prisoners Dilemma cases are much less fantastic than Newcomb problems. Even Prisoners Dilemma cases where we have some confidence that the other party sufficiently resembles us that they will likely (not certainly) make the same choice as us are fairly realistic. So they are somewhat better than Newcomb's original problem for detecting intuitions. But the problem of divergent intuitions still remains. Many people are unsure about what the right thing to do in a Prisoners Dilemma problem is. (We'll come back to this point when we look at game theory.)

So it is worth looking at some cases without that layer of complication. Real life cases are tricky to come by, but for a while some people suggested that the following might be a case. We've known for a long time that smoking causes various cancers. We've known for even longer than that that smoking is correlated with various cancers. For a while there was a hypothesis that smoking did not cause cancer, but was correlated with cancer because there was a common cause. Something, presumably genetic, caused people to (a) have a disposition to smoke, and (b) develop cancer. Crucially, this hypothesis went, smoking did not raise the risk of cancer; whether you got cancer or not was largely due to the genes that led to a desire for smoking.

We now know, by means of various tests, that this isn't true. (For one thing, the reduction in cancer rates among people who give up smoking is truly impressive, and hard to explain on the model that these cancers are all genetic.) But at least at some point in history it was a not entirely crazy hypothesis. Let's assume this hypothesis is actually true (contrary to fact). And let's assume that you (a) want to smoke, other things being equal, and (b) really don't want to get cancer. You don't know whether you have the desire for smoking/disposition to get cancer gene or not? What should you do?

Plausibly, you should smoke. You either have the gene or you don't. If you do, you'll probably get cancer, but you can either get cancer while smoking, or get cancer while not smoking, and since you enjoy smoking, you should smoke. If you don't, you won't get cancer whether you smoke or not, so you should indulge your preference for smoking.

It isn't just philosophers who think this way. At some points (after the smoking/cancer correlation was discovered but before the causal connection was established) various tobacco companies were trying very hard to get evidence for this 'common cause' hypothesis. Presumably the reason they were doing this was because they thought that if it were true, it would be rational for people to smoke more, and hence people would smoke more.

But note that this presumption is true if and only if we use the 'unconditional' version of expected utility theory. To see this, we'll use the following table for the various outcomes.

	Get Cancer	Don't get Cancer
Smoke	1	6
Don't Smoke	0	5

The assumption is that not getting cancer is worth 5 to you, while smoking is worth 1 to you. Now we know that smoking is evidence that you have the cancer gene, and this raises dramatically the chance of you getting cancer. So the (evidential) probability of getting cancer conditional on smoking is, we'll assume, 0.8, while the (evidential) probability of getting cancer conditional on not smoking is, we'll assume, 0.2. And remember this isn't because cancer causes smoking in our example, but rather that there is a common cause of the two. Still, this is enough to make the expected utilities work out as follows.

$$\begin{aligned} Exp(U(Smoke)) &= 0.8 \times 1 + 0.2 \times 6 \\ &= 2 \\ Exp(U(Don't Smoke)) &= 0.2 \times 0 + 0.8 \times 5 \\ &= 4 \end{aligned}$$

And the recommendation is not to smoke, even though smoking dominates. This seems very odd. As it is sometimes put, the recommendation here seems to be a matter of managing the 'news', not managing the outcome. What's bad about smoking is that if you smoke you get some evidence that something bad is going to happen to you. In particular, you get evidence that you have this cancer gene, and that's really bad news to get because dramatically raises the probability of getting cancer. But not smoking doesn't mean that you don't have the gene, it just means that you don't find out that you have the gene. Not smoking looks like a policy of denying yourself good outcomes because you don't want to get bad news. And this doesn't look rational.

So this case has convinced a lot of decision theorists that we shouldn't use conditional probabilities of states when working out the utility of various outcomes. Using conditional probabilities will be good if we want to learn the 'news value' of some choices, but not if we want to learn how useful those choices will be to us.

17.2 Tickle Defence

Not everyone has been convinced by these 'real-life' examples. The counter-argument is that in any realistic case, the gene that leads to smoking has to work by changing our dispositions. So there isn't just a direct causal connection between some genetic material and smoking. Rather, the gene causes a desire to smoke, and the desire to smoke cause the smoking. As it is sometimes put, between the gene and the smoking there has to be something mental, a 'tickle' that leads to the smoking.

Now this is important because we might think that rational agents know their own mental states. Let's assume that for now. So if an agent has the smoking desire they know it, perhaps because this desire has a distinctive phenomenology, a tickle of sorts. And if the agent knows this, then they won't get any extra evidence that they have a desire to smoke from their actual smoking. So the probability of getting cancer given smoking is not higher than the probability of getting cancer given not smoking.

In the case we have in mind, the bad news is probably already here. Once the agent realises that their values are given by the table above, they've already got the bad news. Someone who didn't have the gene wouldn't value smoking more than not smoking. Once the person conditionalises on the fact that that is their value table, the evidence

that they actually smoke is no more evidence. Either way, they are (say) 80% likely to get cancer. So the calculations are really something like this

$$\begin{aligned} Exp(U(Smoke)) &= 0.8 \times 1 + 0.2 \times 6 \\ &= 2 \\ Exp(U(Don't Smoke)) &= 0.8 \times 0 + 0.2 \times 5 \\ &= 1 \end{aligned}$$

And we get the correct answer that in this situation we should smoke. So this isn't a case where the two different equations we've used give different answers. And hence it isn't a reason for using unconditional probabilities rather than conditional probabilities.

There are two common responses to this argument. The first is that it isn't clear that there is always a 'tickle'. The second is that it isn't a requirement of rationality that we know what tickles we have. Let's look at these in turn.

First, it was crucial to this defence that the gene (or whatever) that causes both smoking and cancer causes smoking by causing some particular mental state first. But this isn't a necessary feature of the story. It might be that, say, everyone has the 'tickle' that goes along with wanting to smoke. (Perhaps this desire has some evolutionary advantage. Or, more likely, it might be a result of something that genuinely had evolutionary advantage.) Perhaps what the gene does is to affect how much willpower we have, and hence how likely we are to overcome the desire.

Second, it was also crucial to the defence that it is a requirement of rationality that people know what 'tickles' they have. If this isn't supposed, we can just imagine that our agent is a rational person who is ignorant of their own desires. But this supposition is quite strong. It is generally not a requirement of rationality that we know things about the external world. Some things are just hidden from us, and it isn't a requirement of rationality that we be able to see what is hidden. Similarly, it seems at least possible that some things in our own mind should be hidden. Whether or not you believe in things like subconscious desires, the possibility of them doesn't seem to systematically undermine human rationality.

Note that these two responses dovetail nicely. If we think that the gene works not by producing individual desires, but by modifying quite general standing dispositions like how much willpower we have, it is even more plausible to think that this is not something a rational person will always know about. It is a little odd to think of a person who desires to smoke but doesn't realise that they desire to smoke. It isn't anywhere near as odd to think about a person who has very little willpower but, perhaps because their willpower is rarely tested, doesn't realise that they have low willpower. Unless they are systematically ignoring evidence that they lack willpower, they aren't being clearly irrational.

So it seems there are possible, somewhat realistic, cases where one choice is evidence, to a rational agent, that something bad is likely to happen, even though the choice does not bring about the bad outcome. In such a case using conditional probabilities will lead to avoiding the bad news, rather than producing the best outcomes. And that seems to be irrational.