Chapter 16 Newcomb's Problem

16.1 The Puzzle

In front of you are two boxes, call them A and B. You call see that in box B there is $1000, but you cannot see what is in box A. You have a choice, but not perhaps the one you were expecting. Your first option is to take just box A, whose contents you do not know. Your other option is to take both box A and box B, with the extra $1000.

There is, as you may have guessed, a catch. A demon has predicted whether you will take just one box or take two boxes. The demon is very good at predicting these things – in the past she has made many similar predictions and been right every time. If the demon predicts that you will take both boxes, then she's put nothing in box A. If the demon predicts you will take just one box, she has put $1,000,000 in box A. So the table looks like this.

	Predicts 1 box	Predicts 2 boxes
Take 1 box	$1,000,000	$0
Take 2 boxes	$1,001,000	$1,000

There are interesting arguments for each of the two options here.

The argument for taking just one box is easy. The way the story has been set up, lots of people have taken this challenge before you. Those that have taken 1 box have walked away with a million dollars. Those that have taken both have walked away with a thousand dollars. You'd prefer to be in the first group to being in the second group, so you should take just one box.

The argument for taking both boxes is also easy. Either the demon has put the million in the opaque or she hasn't. If she has, you're better off taking both boxes. That way you'll get $1,001,000 rather than $1,000,000. If she has not, you're better off taking both boxes. That way you'll get $1,000 rather than $0. Either way, you're better off taking both boxes, so you should do that.

Both arguments seem quite strong. The problem is that they lead to incompatible conclusions. So which is correct?

16.2 Two Principles of Decision Theory

The puzzle was first introduced to philosophers by Robert Nozick. And he suggested that the puzzle posed a challenge for the compatibility of two decision theoretic rules. These rules are

- Never choose dominated options
- Maximise expected utility

Nozick argued that if we never chose dominated options, we would choose both boxes. The reason for this is clear enough. If the demon has put $1,000,000 in the opaque box, then it is better to take both boxes, since getting

$1,001,000 is better than getting $1,000,000. And if the demon put nothing in the opaque box, then your choices are $1,000 if you take both boxes, or $0 if you take just the empty box. Either way, you're better off taking both boxes. This is obviously just the standard argument for taking both boxes. But note that however plausible it is as an argument for taking both boxes, it is compelling as an argument that taking both boxes is a dominating option.

To see why Nozick thought that maximising expected utility leads to taking one box, we need to see how he is thinking of the expected utility formula. That formula takes as an input the probability of each state. Nozick's way of approaching things, which was the standard at the time, was to take the expected utility of an action *A* to be given by the following sum

$$Exp(U(A)) = Pr(S_1|A)U(AS_1) + ... + Pr(S_n|A)U(AS_n)$$

Note in particular that we put into this formula the probability of each state *given that A is chosen*. We don't take the unconditional probability of being in that state. These numbers can come quite dramatically apart.

In Newcomb's problem, it is actually quite hard to say what the probability of each state is. (The states here, of course, are just that there is either $1,000,000 in the opaque box or that there is nothing in it.) But what's easy to say is the probability of each state given the choices you make. If you choose both boxes, the probability that there is nothing in the opaque box is very high, and the probability that there is $1,000,000 in it is very low. Conversely, if you choose just the one box, the probability that there is $1,000,000 in it is very high, and the probability that there is nothing in it is very low. Simplifying just a little, we'll say that this high probability is 1, and the low probabiilty is 0. The expected utility of each choice then is

$$\begin{aligned}
Exp(U(\text{Take both boxes})) & \\
&= Pr(\text{Million in opaque box}|\text{Take both boxes})U(\text{Take both boxes and million in opaque box}) \\
&+ Pr(\text{Nothing in opaque box}|\text{Take both boxes})U(\text{Take both boxes and nothing in opaque box}) \\
&= 0 \times 1,001,000 + 1 \times 1,000 \\
&= 1,000 \\
Exp(U(\text{Take one box})) & \\
&= Pr(\text{Million in opaque box}|\text{Take one box})U(\text{Take one box and million in opaque box}) \\
&+ Pr(\text{Nothing in opaque box}|\text{Take one box})U(\text{Take one box and nothing in opaque box}) \\
&= 1 \times 1,000,000 + 0 \times 0 \\
&= 1,000,000
\end{aligned}$$

I've assumed here that the marginal utility of money is constant, so we can measure utility by the size of the numerical prize. That's an idealisation, but hopefully a harmless enough one.

16.3 Bringing Two Principles Together

In earlier chapters we argued that the expected utility rule never led to a conflict with the dominance principle. But here it has led to a conflict. Something seems to have gone badly wrong.

The problem was that we've used two distinct definitions of expected utility in the two arguments. In the version we had used in previous chapters, we presupposed that the probability of the states was independent of the choices that were made. So we didn't talk about $Pr(S_1|A)$ or $Pr(S_1|B)$ or whatever. We simply talked about $Pr(S_1)$.

If you make that assumption, expected utility maximisation does indeed imply dominance. We won't rerun the entire proof here, but let's see how it works in this particular case. Let's say that the probability that there is $1,000,000 in the opaque box is *x*. It won't matter at all what *x* is. And assume that the expected utility of a choice *A* is given by this formula, where we use the unconditional probability of states as inputs.

$$Exp(U(A)) = Pr(S_1)U(AS_1) + ... + Pr(S_n|A)U(AS_n)$$

Applied to our particular case, that would give us the following calculations.

$$\begin{aligned}
&Exp(U(\text{Take both boxes})) \\
&\qquad = Pr(\text{Million in opaque box})U(\text{Take both boxes and million in opaque box}) \\
&\qquad + Pr(\text{Nothing in opaque box})U(\text{Take both boxes and nothing in opaque box}) \\
&\qquad = x \times 1,001,000 + (1-x) \times 1,000 \\
&\qquad = 1,000 + 1,000,000x \\
&Exp(U(\text{Take one box})) \\
&\qquad = Pr(\text{Million in opaque box})U(\text{Take one box and million in opaque box}) \\
&\qquad + Pr(\text{Nothing in opaque box})U(\text{Take one box and nothing in opaque box}) \\
&\qquad = x \times 1,000,000 + (1-x) \times 0 \\
&\qquad = 1,000,000x
\end{aligned}$$

And clearly the expected value of taking both boxes is 1,000 higher than the expected utility of taking just one box. So as long as we don't conditionalise on the act we are performing, there isn't a conflict between the dominance principle and expected utility maximisation.

While that does resolve the mathematical puzzle, it hardly resolves the underlying philosophical problem. Why, we might ask, shouldn't we conditionalise on the actions we are performing? In general, it's a bad idea to throw away information, and the choice that we're about to make is a piece of information. So we might think it should make a difference to the probabilities that we are using.

The best response to this argument, I think, is that it leads to the wrong results in Newcomb's problem, and related problems. But this is a somewhat controversial clam. After all, some people think that taking one box is the right result in Newcomb's problem. And as we saw above, if we conditionalise on our action, then the expected utility of taking one box is higher than the expected utility of taking both. So such theorists will not think that it gives the wrong answer at all. To address this worry, we need to look more closely back at Newcomb's original problem, and its variants.

16.4 Well Meaning Friends

The next few sections are going to involve looking at arguments that we should take both boxes in Newcomb’s problem, or to rejecting arguments that we should only take one box.

The simplest argument is just a dramatisation of the dominance argument. But still, it is a way to see the force of that argument. Imagine that you have a friend who can see into the opaque box. Perhaps the box is clear from behind, and your friend is standing behind the box. Or perhaps your friend has super-powers that let them see into opaque boxes. If your friend was able to give you advice, and has your best interests at heart, they’ll tell you to take both boxes. That’s true whether or not there is a million dollars in the opaque box. Either way, they’ll know that you’re better off taking both boxes.

Of course, there are lots of cases where a friend with more knowledge than you and your interests at heart will give you advice that is different to what you might intuitively think is correct. Imagine that I have just tossed a biased coin that has an 80% chance of landing heads. The coin has landed, but neither of us can see how it has landed. I offer you a choice between a bet that pays $1 if it landed heads, and a bet that pays $1 if it landed tails. Since heads is more likely, it seems you should take the bet on heads. But if the coin has landed tails, then a well meaning and well informed friend will tell you that you should bet on tails.

But that case is somewhat different to the friend in Newcomb’s problem. The point here is that you know what the friend will tell you. And plausibly, whenever you know what advice a friend will give you, you should follow that advice. Even in the coin-flip case, if you knew that your friend would tell you to bet on tails, it would be smart to bet on tails. After all, knowing that your friend would give you that advice would be equivalent to knowing that the coin landed tails. And if you knew the coin landed tails, then whatever arguments you could come up with concerning chances of landing tails would be irrelevant. It did land tails, so that’s what you should bet on.

There is another way to dramatise the dominance argument. Imagine that after the boxes are opened, i.e. after you know which state you are in, you are given a chance to revise your choice if you pay $500. If you take just one box, then whatever is in the opaque box, this will be a worthwhile switch to make. It will either take you from $0 to $500, or from $1,000,000 to $1,000,500. And once the box is open, there isn’t even an intuition that you should worry about how the box got filled. So you should make the switch.

But it seems plausible in general that if right now you’ve got a chance to do X, and you know that if you don’t do X now you’ll certainly pay good money to do X later, and you know that when you do that you’ll be acting perfectly rationally, then you should simply do X. After all, you’ll get the same result whether you do X now or later, you’ll simply not have to pay the ‘late fee’ for taking X any later. More relevantly to our case, if you would switch to X once the facts were known, even if doing so required paying a fee, then it seems plausible that you should simply do X now. It doesn’t seem that including the option of switching after the boxes are revealed changes anything about what you should do before the boxes are revealed, after all.

Ultimately, I’m not sure that either of the arguments I gave here, either the well meaning friend argument or the switching argument, are any more powerful than the dominance argument. Both of them are just ways of dramatising the dominance argument. And someone who thinks that you should take just one box is, by definition, someone who isn’t moved by the dominance argument. In the next set of notes we’ll look at other arguments for taking both boxes.