

Newcomb's Paradox and the Sorites

11th May 2007

1 The Game

Newcomb's Paradox is so-called because it proposes a simple game for which two apparently equally sensible yet contradictory strategies exist. Simply stated, it goes like this. Imagine there is a game in which each contestant is presented with two boxes. Box A always contains £1,000; the contents of Box B are hidden. You, the contestant, are presented with the choice of taking both Box A and Box B, or of taking only Box B.

The catch is this: the game is operated by a shady character called the Predictor, who seems uncannily good at guessing which choice you will make. If it thinks you'll take both boxes, it leaves Box B empty. If it thinks you'll take Box B only, it instead puts £1,000,000 there. The Predictor is right ninety-nine times out of a hundred.

Nobody knows how the Predictor makes predictions, whether it's by lucky guesses, using a computer simulation or psychological manipulation. Some think such accuracy must involve ESP or time travel. One thing is certain: the contents of the boxes are fixed before you make your decision. Perhaps they're displayed on television screens all over the country although you, of course, can't see them. The question is, should you take both boxes or only Box B when it comes to your turn to play?

2 Expected Utility

Initially it seems obvious that you should take only Box B. Imagine that the Predictor is very accurate, and that you have seen many games like this played before. Then in the vast majority of cases taking Box B yielded £1,000,000 while those who took both boxes walked out with a mere £1,000. If you'd seen that for yourself, wouldn't you choose Box B alone when it came to your turn to play? Why try to buck the trend when most people who did came out worse off?

You can firm up this intuition by employing the idea of 'expected utility'. Essentially, you say that choosing Box B yields £1,000,000 99% of the time, so if you played it a hundred times you'd get £99,000,000 – the one time the Predictor got it wrong, you'd get nothing. Dividing by a hundred – since you

only really get one chance to play – gives a utility of £990,000 for opening Box B.

Now look at the other option – taking both boxes. This time you get £99,000 for the ninety-nine times the Predictor is right, plus £1,001,000 for the one time the Predictor is wrong, and dividing by a hundred again gives us a utility of £11,000 for opening both boxes.

By this line of reasoning, the utility of choosing only Box B is *far* more than that of choosing both boxes. You'd be insane to choose both boxes given the success rate the Predictor has had in the past.

3 Dominance

Yet there's another thing to consider. Imagine we change the game now so that you can see into the boxes. What would you see? Well, Box A always has £1,000 in it. Box B would have either £1,000,000 or nothing. What decision would you make then? Of course, you'd always take both boxes. Why wouldn't you? The contents of the boxes are already decided, and so you might as well walk out with whatever's on the table. If that's what you'd always do if you could see into the boxes then what, superstition aside, would cause you to do anything different in the real game, where the boxes are opaque?

This is known as the 'dominance' strategy. The proponent of expected utility makes £1,000,000 in the best case and nothing in the worst. The advocate of dominance makes £1,001,000 in the best case and £1,000 in the worst.

What's changed? Not the contents of the boxes, since they're fixed when you walk into the room, so changing your strategy can't affect those – can it?. Well, in fact the proponent of the utility argument thinks that *if* you choose Box B only, *then* it's far more likely to contain the money than if you choose both. Based on past experience the success and failure scenarios are by no means equally likely.

In other words, the utilitarian has built into her calculations the notion that *the choice you make affects the contents of the box*. Although based on empirical evidence (the observed success rate) it's a very disturbing idea, since we don't normally think that an event happening now (choosing Box B) can determine what happened in the past (whether the Predictor left Box B empty).

Can such 'backwards causation' really happen? Some people think it can. If the Predictor travels forward in time to watch you make your choice, and then goes back in time and fills the boxes, then the utility argument makes perfect sense. But what if we could rule out this possibility?

4 A Strategic Sliding Scale

Part of the problem as it's stated is this high success rate of 99%. It just seems unbelievable, unless something fishy – ESP, time travel or outright cheating – is going on. What changes if we reduce the success rate? *If* some sort of backward

causation is going on then utility outperforms the dominance strategy until you get very close to 50% accuracy, so nothing really ought to change, but I think something does.

Say the Predictor has been only 60% accurate so far, and not very many games have been played – maybe only a few dozen. Now, the utility strategy is still better than dominance if there’s a causal relationship between your action and the contents of Box B. But we’re really no longer inclined to think there is. A success rate of 60% over a few trials might very well just be luck. You’d have to a rather eager exponent of time travel or ESP to believe that such exotic explanations are required to account for this new situation.

As such, you would probably adopt the dominance position, since the utility argument is no longer plausible. If you don’t accept that backwards causation is likely in this situation then to adopt the utility strategy would be simply to leave £1,000 on the table when you played the game. The contents of Box B are already present (or absent) and your choice has no impact on them.

It seems as if there’s a sliding scale here. At one extreme, the Predictor is extremely accurate, and has proved its accuracy over a very large number of games. Under these circumstances, it is not entirely unreasonable to assume that the choice you make affects the contents of the box even if, as a sceptic about both ESP and time travel, you simply believe that the game is rigged.

At the extreme, this situation is equivalent to a variation on the game, in which you are told openly that you will *definitely* get £1,000,000 if you choose only Box B, or £1,000 if you open both. In this case, opening Box B only, based on its expected utility, is clearly the best strategy.

At the other end of the scale, the Predictor has not got such a good track record; correct predictions occur only slightly more than 50% of the time, and not many games have been played. In this case, we’re not inclined to believe that our choice can affect the contents of the boxes. Here the dominance strategy is obviously superior.

Now, take that (lower) end of the scale and imagine increasing the success rate. Gradually we become more likely to favour the utility argument, and to suspect some causal link between our choice and the contents of Box B. Eventually, as the success rate comes close to 100%, we would be foolish to cling to the dominance argument.

5 Sliding the Scale

The problem with all this, of course, is that there’s no obvious cutover point between believing that the game is fair and believing that some sort of causation is at work. Instead, we shift more or less willingly from the dominance position, in which things are normal and predictions about our freely-taken decisions are right about half the time, and the utility position, in which we accept that our choices are affecting the game.

This happens somewhere between 50% success rate and 100%. But where? Halfway – that is, at 75%? I think it depends on your willingness to flip into

the utility position. Most people are likely to strongly resist the idea that their actions in the present can affect events in the past (assuming the possibility of cheating is minimised). They'll find more or less rational explanations for a success rate of 80%, say, that have nothing to do with a causal link between the choice and the contents of the box. Nevertheless, as the success rate grows so does the difficulty of maintaining such a position.

On the other hand, many people are susceptible to superstitions far more easily than a figure of 75% would suggest. Just the presence of a Predictor, even if its success rate were close to 50%, would doubtless be enough to convince some people to open Box B only because that was what was wanted of them, and they'd seen others who made the same choice rewarded.

6 The Sorites Problem

What's the rational choice? Where on the strategy scale would the truly rational person change strategies? This looks like a classic instance of the sorites problem. The term 'sorites' is derived from the Greek word for 'heap'. The story goes that a merchant offered for sale a heap of ground flour. One heap, ten dinarii. Soon he had a buyer, and they entered into a binding transaction. 'Before you take it,' said the merchant, 'do you agree that removing one grain from this heap doesn't stop it being a heap?'

'Of course,' the naïve buyer replies. The merchant removes one grain, and repeats his question. So the afternoon passes, until there is very little flour left on the table and everyone else has gone home. The point is that there's no single point at which the buyer could have said 'Stop there; up to now, removing one grain didn't change the heap into a non-heap, but the next grain would'. There's a sort of fuzzy area in the middle; entering it, like flying into a cloud, is a gradual process, but when you come out the other side you find yourself somewhere different. You had a heap of flour but you no longer do; the philosophical problem is that it's not possible to say exactly when this reversal happened.

Instead what we have is effectively a continuum, with a definite heap at one end and a definite non-heap at the other. Somewhere in between we think there must be a cutoff point, but in fact the whole middle section of the continuum is obscured by a sort of cloud; when we enter the cloud we have a heap, and when we leave it we don't but, the a volunteer from the audience, we can't see exactly when the magician did the trick.

Now, without getting into the sorites problem and the accompanying idea of semantic vagueness, which have been a topic of renewed philosophical debate in recent decades, I think we can say this is a helpful way of modelling the 'strategy flip'. Indeed, it might be a useful way to think of 'changing you mind' in general, in cases where a lot of small pieces of evidence pile up – the Predictor continues to be right more often than wrong – and at some point you find that you believe something different from what you believed before.

One very simple way of dealing with the sorites is using a bit of probability.

At one end we have a heap with probability 1 – that is, we definitely have a heap. At the other, we have a heap with probability 0 – we definitely don't have a heap. In between we let the probability slide smoothly between the two values. We don't mean literally that this is a 'probability' of the flour forming a heap, but that it models whatever we think explains vagueness; philosophers are divided as to whether it's really semantic or epistemic or even ontological. That doesn't matter for us; we just want to attached a number to it that describes how bad it is. The closer to 0.5 or 50% (which means the same thing) the more vague we are.

7 A Rational but Not Very Good Strategy for Multiple Games

Now imagine yet another change in the nature of the problem. In this case, you are the only player, and you will play game after game. We'll assume that the Predictor isn't going to learn from how you play; perhaps you'll play a thousand times, and in advance the Predictor makes a thousand predications.

What's the best strategy? Well, at first you don't know anything, so assume the simplest situation, which is that the Predictor is no more than a Guesser. Hence you'll play the dominance strategy. But with each game, be sure to recalculate the Predictor's success rate.

Let's also assume that the success rate is always better than, or close to, 50%. If it drops well below this then the Predictor is, of course, making very good predictions but not the expected ones, and the utility strategy comes into play again in reverse. You should always take both boxes in such a situation, for you'll usually catch the Predictor out and take home £1,001,000. Proponents of either strategy can at least agree on that.

Now, in any subsequent game, you might play like this. If the Predictor's success rate is less than 50%, play the dominance strategy. Otherwise, play the dominance strategy with a probability of

$$2(100 - s)\%$$

where s is the success rate. So if the success rate is 75%, you'll play the dominance strategy about half the time and the strategy based on expected utility otherwise. Of course, if the success rate climbs close to 100% then you'll almost always play based on expected utility, whereas if it stays around 50% you'll stick with dominance on the assumption that there's no backwards causality in the mix.

The problem is, this strategy doesn't net you the most money. As we've already said, if strong backward causation actually is at work then expected utility is the way to go, whereas otherwise dominance will be the winning strategy. The problem with Newcomb's Paradox is that *we just don't know* how the Predictor's predictions are made. I think this is the nub of the problem, and the thing that makes it philosophically suggestive.

8 Leaps of Faith

The strategy outlined above enables the player to learn to what degree the game is likely to have an element of backward causation, and to profit from it without making a one-off ‘leap of faith’ at a more or less arbitrary point. There’s a problem, though; if backward causality is at work even a little – say, enough to give a 60% success rate – then it’s much better to play the expected utility strategy. In real life, someone using the strategy just outlined might very well deviate from it if they saw a high success rate emerging, and abandon it entirely well below 100%.

Imagine you belong to a community that sacrifices one percent of its crops every year to the gods, expecting in return a good harvest next year. Clearly if you stop doing that you’ll instantly increase crop yields by the amount that’s thrown away. But imagine, too, that you’ve had fifteen good harvests in the last twenty years while another community on the other side of the river doesn’t sacrifice to the gods and has had only ten.

This situation is partially parallel to Newcomb’s Paradox. We, if we are scientifically-minded or even if we simply don’t share their religious system, will look for other reasons for the disparity. Nevertheless, it’s hard to say that they should flip strategies when the one they’re using seems to be working better than the one across the river. It’s ‘partially’ parallel because, unlike the artificial situation presented in the game, this is a real-world situation in which many complex factors are likely to interact. As a result, we would ask for a very high disparity between the success rates of the two communities’ agriculture, and an absence of any other explanations, before we accepted that the sacrifice really was doing some good.

There’s a probabilistic version of Pascal’s Wager going on here, since the two strategies ‘ought to’ perform about the same (a 1% difference, which is the sacrifice) but the reward if the utility position is correct is significant. Hence that position may be adopted on the basis of quite weak evidence simply because it’s cheap and it might work.

We often see a similar line of reasoning when talking about more or less faith-based activities. For example, even strong sceptics sometimes accept the use of homeopathy as a complement to mainstream medicine since it can really do very little harm if it doesn’t work and – who knows? – it might do some good. Of course there is a cost, either to the patient or the state, incurred by the utility argument, just like the 1% sacrifice. The question is, are the purported benefits worth it?

9 Another Dimension on the Scale

The situation portrayed in the original game involves a big disparity between winning and losing. Whichever strategy you play, you stand to gain about a million pounds if you win, in contrast to which the thousand pounds you walk away with if you lose seems unimpressive.

What if the two figures were closer together? Well, too close and the game is no fun. Say Box A always contains £1,000,000, and Box B contains either £1,000,000 or nothing. Well, now you're looking at a choice of a definite £1,000,000, with a small chance of £2,000,000 (by picking up both boxes) or only a good chance of £1,000,000 (by choosing only Box B). The dominance strategy is obviously the winner here even if the success rate of the Predictor is 100%.

As we decrease the amount in Box A, and assuming the success rate is fairly high, we're likely to start favouring utility well before it gets to £1,000. Of course, at the other end of the scale it contains nothing, so you may as well choose Box B every time however lousy a Predictor you have.

I think, then, that the extent to which we're willing to 'play along' with the Predictor is determined in part by what's at stake. We should only be willing to make the 'leap of faith' if the two things are in agreement: the success rate, which corresponds to our estimated likelihood that the faith position is true, and the stakes. We always stand to gain from the dominance strategy *if it's right*, but how much we have to gain has to be balanced against how likely we think it is that it really is right.

What makes matters worse is that often the stakes are not clearly-defined. If homeopathy works, what does that mean for a particular patient? That their illness will be instantly and fully cured? Or cured over a long period? Or mitigated? Or will it address only the symptoms? To what extent? There's obviously yet another continuum here from full, immediate recovery all the way to no effect at all. To say that it 'works' isn't sufficient; we need to know what will happen if the present case is one of the $s\%$ of cases in which it actually *does* work.

10 The Whole Picture

We now have quite a complicated structure for operating in Newcomb's Paradox-like situations. First, we have a continuum from 50% to 100% that indicates our willingness to flip strategies in a zero-sum version of the game. This is extended into a plane by considering another continuum of expected utility. In the original version we were dealing with sums of money, so this was known for certain, but at the other extreme, which is more usual in the real world, we have a cloud of uncertainty around it.

See Figure 1. Your situation will be a point in the square. Running left to right is what you have to gain by playing the dominant strategy *if* the assumption of no causal relationship is correct. Running bottom to top is the success rate. As the success rate increases, the stakes need to be higher and higher before you'll play the dominance strategy.

In the middle is a cloud of uncertainty runs along the dividing line, representing the sorites-like problem of a tiny change in either variable leading to a complete switch of strategy. It seems to me that we're often stuck in this cloud in cases where we must decide between two contradictory strategies based on

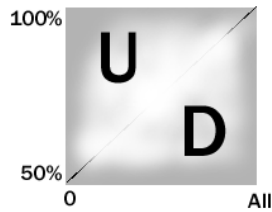


Figure 1: A space of Newcomb's Paradoxes

insufficient information, and in particular when the historical success rate of one seems unwarranted by the available evidence.

11 Reductions and Generalisations

According to this model, Newcomb's Paradox is characterised by a pair of orthogonal (ie independent) variables that are subject to sorites conditions. In fact, though, this is suggestive of much more. First, any situation that involves such a pair of variables is equivalent to Newcomb's Paradox. That means that if we have a strategy for dealing with Newcomb's Paradox then we potentially have a strategy for dealing with other, equivalent problems.

Second, we can reduce Newcomb's Paradox to a pair of sorites problems, so we can deal with it using strategies for dealing with them. The sorites problem is very much alive and well in philosophical literature, so this doesn't give us an easy answer, but it's definitely a better-understood and simpler problem than Newcomb is.

Third, it leads us to imagine the kinds of structure that might emerge in the case of more than two independent variables. Higher-dimensional paradoxes are created by adding more independent variables each of which is subject to a cloud of sorites-induced vagueness. It seems likely that real-life situations that seem equivalent to Newcomb actually have many more dimensions; it's just a question of how sophisticated a model you need.

References

- [1] Gardner, Martin, *Knotted Doughnuts and Other Mathematical Entertainments*, W H Freeman, New York, 1986