Newcomb's problem, expected utility, and dominance

PHIL 20229
Jeff Speaks

March 25, 2008

1 Newcomb's problem 1
2 Maximizing expected utility 2
3 Dominance rules 3
4 How does the Predictor know? 3

1 Newcomb's problem

Let's focus on the following version of Newcomb's problem:

> There are two boxes before you, Box A and Box B. You have a choice as to whether you can take only the contents of Box B, or can take the contents of Box A and Box B. The Predictor has placed $1000 in Box A. If the Predictor predicts that you will take only Box B, he has placed $1,000,000 in Box B. If the Predictor predicts that you will take the contents of both boxes, he has placed nothing in Box B. You've observed that, in the past, the Predictor is right every time. If your aim is to maximize your money, should you choose to take the contents of Box B alone, or the contents of both boxes?

This case is a problem because there seem to be intuitively compelling arguments for the rationality of two conflicting courses of action: selecting one box (one-boxing) or selecting two boxes (two-boxing). The arguments can be stated informally as follows:

Argument for one-boxing: You have no reason to think that your case is different than anyone else's. (In Sainsbury's version of the case, the Predictor has invariably predicted your actions, not the actions of people deciding between one- and two-boxing.) So you should think that in your case as in the others, one-boxing will lead to a higher payout than two-boxing. Since the aim of the game is to maximize your money, you ought to one-box.

Argument for two-boxing: The money is already in the boxes. If you one-box, you will get whatever is in box B, whereas if you two-box you will get this plus the $1000 that is in box A. So you know in advance that two-boxing will give you $1000 more than one-boxing, and hence you should two-box.

Each of these arguments employs a plausible principle about rational decision making. One of the interesting aspects of Newcomb's problem is that it seems to be a place where these two principles come into conflict.

2 Maximizing expected utility

The first principle — which plays a role in the argument for one-boxing — is the principle that it is rational to act so that you maximize expected utility.

You can think of utility as benefit. What counts as a benefit in a given case might depend on what you value. In the example of Newcomb's problem, we simplify by just restricting the relevant utility to monetary reward.

Measures of expected utility weigh the utility of various outcomes of a given course of action against the probability of the various outcomes in order to tell you how much utility you should expect from that course of action. This is, intuitively, the way that we think (or ought to think) about bets and gambling. Suppose that a slot machine has a very high maximum payoff; this counts in favor of playing that machine. But this advantage may be negated if the slot has an extremely low probability of rewarding a player with the maximum payoff.

To calculate the expected utility of a course of action, you assign utility measures to the various possible outcomes, and assign probabilities to those outcomes. The expected utility is the sum of the utilities of the outcomes times their probabilities.

Given this definition, then it seems straightforward that in situations like the one described above, one should always act so as to maximize expected utility. But then it seems straightforward that one should be a one-boxer. For we can calculate the expected utilities of one-boxing and two-boxing as follows.

Expected utility of one-boxing:

Outcome 1: $1,000,000. Probability: 100%

Outcome 2: $0. Probability: 0%

Expected utility: $1,000,000.

Expected utility of two-boxing:

Outcome 1: $1,001,000. Probability: 0%

Outcome 2: $1000. Probability: 100%

Expected utility: $1000.

One-boxing has a *much* higher expected utility than two-boxing — 1000 times higher. The case in favor of one-boxing therefore seems fairly straightforward.

Suppose that, even though the Predictor has always been right in the past, you think that there is some chance that he will be wrong in your case. Suppose that you think that, despite getting every case he's ever considered right, there's a 10% chance that he will go wrong for the first time in your case. Would this make it rational to two-box?

3 Dominance rules

The problem is that there seems to be an equally strong argument in favor of two-boxing. Rather than relying on the principle that we should act so as to maximize expected utility, this argument relies on the *dominance principle*:

> Suppose that you are choosing between two actions, act 1 and act 2. It is always rational to do act 2 if the following is the case: whatever happens, doing act 2 will never make you worse off than doing act 1; and in some cases, doing act 2 will make you better off than doing act 1.

Like the principle that we should act so as to maximize expected utility, the dominance principle seems obviously correct. To illustrate this, suppose that you are offered two bets, one of which has a maximum payoff twice that of the other, while everything else about the bets is the same. Isn't it just obvious that you should take the bet with the higher maximum payoff? And isn't this just because that bet, in the above sense, dominates the other?

But then it seems clear that we should be two-boxers, since two-boxing seems to dominate one-boxing, as the following way of thinking about the outcomes of the case shows:

	The Predictor has placed $1,000,000 in Box B	The Predictor has placed nothing in Box B
Two-box	$1,001,000	$1000
One-box	$1,000,000	$0

There are two possible situations, one in which the Predictor has put the cash in Box B, and one in which he has not. In either situation, you are better off two-boxing. In other words, no matter what the Predictor has done, you are better off two-boxing. So it seems fairly clear that it is rational to be a two-boxer.

4 How does the Predictor know?

One wonders, when thinking about the problem, how the Predictor can be so good at telling what people will do. Distinguishing two different answers to this question can help us by distinguishing three importantly different versions of the case.

The backward causation version of the case. In this case, one-boxing is clearly the correct choice. Is this a counterexample to the dominance principle? Why does the dominance principle fail in this version of the scenario?

The no-backward-causation version. Suppose that we disallow backward causation: the Predictor places money in the boxes prior to your making your choice, and cannot tamper with the amounts after you made the choice. Then presumably the Predictor assigns amounts to the two boxes based on an intricate knowledge of your psychology and habits, the state of your brain when you enter the room, etc. What is it rational to do in this version of the case?

Suppose that we think that in the no-backward causation version of the case, two-boxing is rational. Then the principle that we should always act so as to maximize expected utility seems to be false. Can we give any explanation of why the principle fails in this case? Does this suggest any way of restricting the principle so that it might be true?