Topic 2: Newcomb's Problem

Contents

The Problem

Newcomb's Problem

Imagine that you are led into a room, and presented with two boxes: a large one and a small one. You know that the small box contains a thousand dollars. You're not sure what the large one contains, but you know that it either contains a million dollars or is completely empty.

You are offered two choices:

- One-Box
  Keep the contents of the large box, but leave the contents of the small box behind.
- Two-Box
  Keep the contents of both boxes.

How should you proceed? Should you one-box or should you two-box?

The answer seems totally obvious. You should take both boxes!

How could you possibly benefit from leaving a thousand dollars behind? Whether or not the large box contains a million dollars, you'll end up with more money if you take the small box as well!

The Twist

Wait! There's a twist... Let me tell you what happened before you entered the room.

A couple of weeks ago, a personality expert was summoned, and was handed as much information about you as could be gathered. The expert was then asked to determine, on the basis of that information, whether you are a one-boxer (i.e. the kind of person who would only take the large box), or two-boxer (i.e. the kind of person who would take both boxes). If the expert concluded that you are a one-boxer, the large box was filled with a million dollars. If she concluded that you are a two-boxer, the large box was left empty.

In other words:

```
##    Veredict of Expert Contents of Large Box Contents of Small Box
## 1 You are a One-Boxer            $1,000,000               $1,000
## 2 You are a Two-Boxer                    $0               $1,000
```

Both boxes have been sealed since last night, and will remain sealed until you make your decision. So if the large box was filled with a million dollars last night, it will continue to hold a million dollars, regardless of what you decide. And if it was left empty last night, it will remain empty regardless of what you decide.

One final point: the expert is known to be highly reliable. She has participated in many thousands of experiments of this kind, and has made accurate predictions 99% of the time. There is nothing special about your case, so you should think that the expert is 99% likely to correctly predict whether you will one-box or two-box.

How should you proceed now that you know about the procedure that was used to fill the boxes? Is it still obvious that you should two-box?

(Keep in mind that the expert knows that you'll be told about how the experiment works, and about how method that was used to decide how to fill the boxes. So she knows that you'll be engaging in just the kind of reasoning that you are engaging in right now!)

The Predicament

The fact that the expert is 99% likely to correctly predict whether you will one-box or two-box gives you the following two pieces of information:

1. if you decide to take both boxes, it is almost certain (99%) that the large box will be empty;
2. if you decide to leave the small box behind, it is almost certain (99%) that the large box will contain a million dollars.

Before learning about the expert, it seemed clear that two-boxing was the right thing to do. But now you know that if you one-box you are almost certain to end up with a million dollars, and that if you two-box you are almost certain to end up with just a thousand. Should you be a one-boxer after all? What to do?

This problem is often referred to as 'Newcomb's Problem' after Lawrence Livermore National Laboratory physicist William Newcomb. It captured the imagination of the philosophical community in the last third of the twentieth century, largely thanks to the writings of Harvard philosopher Robert Nozick and the legendary Scientific American columnist Martin Gardner.

VIDEO REVIEW: The Problem

Could a Newcomb Predictor Really Exist?

When someone is first told about Newcomb's Problem, they often worry about whether there could really be an expert of the kind the story requires.

It is certainly *logically possible* for there to be such an expert. Consider, for example, a super-scientist who creates a molecule-by-molecule duplicate of your body, and makes her prediction by exposing the duplicate to the Newcomb setup and observing the results. As long as you and your duplicate are subjected to identical stimuli, the two of you are almost certain to carry out the same reasoning, and reach the same decision. So there is no reason such a predictor couldn't be 99% reliable, or more.

Of course, none of this shows that perfect (or near-perfect) predictions are possible in practice. But for our purposes it doesn't matter if the Newcomb scenario could be carried out in practice or not. We will be using the Newcomb scenario as a *thought experiment* to help us understand what rational decision-making is all about. So all we really need is for Newcomb scenarios to be logically possible.

Exercise

Even if real-life predictors are not necessary for our purposes, one might be curious to know whether accurate predictors are possible in practice. Suppose someone happens to have a twin. Maybe one could get somewhat reliable predictions about whether she is a one-boxer or a two-boxer by observing whether her twin one-boxes or two-boxes when presented with a version of the Newcomb experiment. Can you come up with a method that could be used to generate Newcomb predictions in practice, even if the subject doesn't have a twin?

- Here is one possible method. (I'm sure there are many others.) - The predictor asks the subject to participate in a Newcomb scenario and observes the results. The subject is then asked to take an amnesiac drug, and repeats the experiment. For the repeated experiment, the predictor predicts that the subject will make the same decision as before.

Maximizing Expected Value

In the previous section I tried to give you an intuitive sense of why the Newcomb scenario might seem puzzling. Now I will give you some mathematical machinery that will help make the puzzle a little more precise.

Suppose that you have a choice between different possible actions. According to standard decision theory, you should adhere to the following principle:

- **Principle of Expected Value Maximization**
  Select the action with the highest expected value.

What is expected value? It is a measure of how desirable the world is expected to be, on the assumption that the action is performed. (More precisely: the expected value of an action is the weighted average of the value of the possible outcomes of that action, with weights given by the probability of the outcome given the action.)

Don't worry... we'll talk more about expected value below!

An Example

Suppose a coin will be tossed. You know that the coin is fair. (You know, in other words, that it has a 0.5 chance of landing Heads and a 0.5 chance of landing Tails.) You are offered a choice between the following two bets:

- **Bet 1** (B_1)
  You get $1000 if the coin lands Heads, and are forced to pay $200 if the coin lands Tails.
- **Bet 2** (B_2)
  You get $100 if the coin lands Heads, and get $50 if the coin lands Tails.

To keep things simple, I will assume that you value only money. (More precisely: I will assume that the degree to which you value a given outcome corresponds to the amount of money you end up with. So, for example, you value an outcome in which you end up with $1000 to degree 1000, and you value an outcome in which you end up paying $200 to degree -200.)

The **expected** value of a bet is the weighted average of the values of the different possible outcomes (with weights given by the probability that the outcome will occur, given that the action is performed).

Consider Bet 1. If you decide to take that bet, there are two possible outcomes:

- H
  The coin lands Heads
- T
  The coin lands Tails

The expected value of Bet 1 is then calculated as follows:

$$EV(B_1) = v(H\ B_1)\text{ů}p(H|B_1) + v(T\ B_1)\text{ů}p(T|B_1)$$

Where the components of the equation are to be understood as follows:

- $v(H\ B_1)$ is the degree to which you value a situation in which the coin lands Heads, having taken Bet 1.
  - Since you value only money, and since you'll receive $1000 in such a situation, $v(H\ B_1) = 1000$.
- $p(H|B1)$ is the probability that you assign to the coin landing Heads, given that you take Bet 1.
  - Since we are dealing with a fair coin, $p(H|B_1) = 50\% = 0.5$.
- $v(T\ B_1)$ is the degree to which you value a situation in which the coin lands Tails, having taken Bet 1.
  - Since you value only money, and since you'll be forced to pay $200 in such a situation, $v(H\ B_1) = -200$.
- $p(T|B1)$ is the probability that you assign to the coin landing Tails, given that you take Bet 1.
  - Since we are dealing with a fair coin, $p(T|B_1) = 50\% = 0.5$.

Filling in the equation above with these values, we get the following result:

$$EV(B_1) = 1000\text{ů}0.5 + (-200)\text{ů}0.5 = 400$$

Upshot: the expected value of Bet 1 is 400.

Exercises

1. Calculate the expected value of Bet 2.

`75`

$$EV(B_1) = v(H\ B_1)\mathring{u}p(H|B_1) + v(T\ B_1)\mathring{u}p(T|B_1) = 100\mathring{u}0.5 + 50\mathring{u}0.5 = 75$$

2. If your options are Bet 1 and Bet 2, what does the Principle of Expected Value Maximization say you should choose?

- `Bet 1`
  - We noted above that the expected value of Bet 1 is 400, and the result of Exercise 1 is that the expected value of Bet 2 is 75. Since 400 is greater than 75, the Principle of Expected Value Maximization asks us to select Bet 1.
- Bet 2

VIDEO REVIEW: An Expected Value Calculation

The General Case

In the general case, the expected value of an action A is defined as follows:

$$EV(A) = (v(O_1A)\mathring{u}p(O_1|A)) + (v(O_2A)\mathring{u}p(O_2|A)) + \ldots + (v(O_iA)\mathring{u}p(O_i|A))$$

where $O_1, O_2, \ldots, O_i$ is an exhaustive list of possible outcomes, any two of which are mutually exclusive.

Equivalently:

$$EV(A) = \sum_{i \in I}(v(O_iA)\mathring{u}p(O_i|A))$$

where the O_i (for i in I) constitute an exhaustive list of possible outcomes, any two of which are mutually exclusive.

Back to Newcomb

What are the expected values of one-boxing and two-boxing?

There are two actions we're considering:

- **One-Box** ($1B$)
  Take the contents of the large box only.
- **Two-Box** ($2B$)
  Take the contents of both boxes.

And, there are two possible outcomes:

- F
  the large box is full
- E
  the large box is empty

The expected value of one-boxing can then calculated as follows:

$$EV(1B) = v(F\ 1B)\mathring{u}p(F|1B) + v(E\ 1B)\mathring{u}p(E|1B)$$

If you one-box and the large box is full, you get \$1,000,000. So, on the assumption that you value only money, $v(F\ 1B) = 1,000,000$. If you one-box and the large box is empty, you get nothing. So, on the assumption that you value only money, $v(E\ 1B) = 0$.

We therefore have:

$$EV(1B) = 1,000,000\mathring{u}p(F|1B) + 0\mathring{u}p(E|1B) = 1,000,000\mathring{u}p(F|1B)$$

What is $p(F|1B)$? In other words: what is the probability that the large box will be full, given that you one-box? If we assume that the predictor is 99% accurate, the answer is 99% = 0.99. So:

$$EV(1B) = 1,000,000\mathring{u}0.99 = 990,000$$

So the expected value of one-boxing is 990,000. What about the expected value of two-boxing? In this case we get the following results:

$$EV(2B) = v(F\ 2B)\mathring{u}p(F|2B) + v(E\ 2B)\mathring{u}p(E|2B) = 1,001,000\mathring{u}0.01 + 1,000\mathring{u}0.99 = 10,010 + 990 = 11,000$$

So whereas the expected value of one-boxing is close to a million (990,000), the expected value of two-boxing is close to ten thousand (11,000).

Upshot: the Principle of Expected Utility Maximization tells us that we should one-box.

(Is that the right result? We'll turn to that topic in the next section.)

VIDEO REVIEW: Calculating EV in the Newcomb Problem

Exercises

1. Assume the predictor has an 80% chance of making the right prediction. What is the expected value of one-boxing?

```
800000
```

When $p = 80\%$, the expected value of one-boxing is

$$(1,000,000 \times 0.8) + (0 \times 0.2) = 800,000$$

What is the expected value of two-boxing?

```
201000
```

The expected value of two-boxing is

$$(1,001,000 \times 0.2) + (1,000 \times 0.8) = 200,200 + 800 = 201,000$$

2. How low can you go? How accurate does the predictor be in order for the Principle of Expected Value Maximization to entail that one should one-box? Enter the smallest number x such that, as long as the predictor has an accuracy of more than x, the Principle of Expected Value Maximization entails that one should one-box. (So, for example, if you think the Principle will recommend one-boxing as long as the predictor is more than 73.22% accurate, enter .7322)

`.5005`

If p is the probability that the expert's prediction is accurate, the expected value of one-boxing is

$$(1,000,000 \times p + (0 \times (1-p))$$

and the expected value of two-boxing is

$$(1,001,000 \times (1-p)) + (1,000 \times p)$$

These two expected values are equal when $p = 0.5005$ — i.e., 50.05%. So as long as p is greater than 50.05%, the expected value of one-boxing will be greater than the expected value of two-boxing.

This means that as long as the predictor is at least 50.05% accurate, the Principle of Expected Value Maximization entails that one should be a one-boxer.

(That is pretty remarkable. Even if experts that are accurate 99% of the time exist only in science fiction, it is not hard to find a predictor that is accurate 50.05% of the time. I myself have an 80% success rate when performing this experiment on my students!)

3. Consider a variant of the Newcomb case which works as follows. You can choose either the large box or the small box, but not both. If the predictor predicts that you would choose the large box, then she left the large box empty, and placed $100 in the small box. If the predictor predicted that you will choose the small box, then she left the small box empty, and placed $1000 in the large box. According to the Principle of Expected Value Maximization, which of the two boxes should you choose? (Assume the predictor is not perfectly accurate.)

- `Large box`
  - The expected value of choosing the large box is

    $$(1000 \times (1-p)) + (0 \times p) = 1000 \times (1-p)$$

    The expected value of choosing the small box is

    $$(100 \times (1-p)) + (0 \times p) = 100 \times (1-p)$$

    So the Principle of Expected Value Maximization entails that one should choose the large box (for every case except $p = 1$; in that case, both options have expected value 0, so the Principle of Expected Value Maximization doesn't entail that you ought to choose one option over the other.).
- Small box
- Either
- Problem is ill-defined

---

3. In Defense of Two-Boxing

In Defense of Two-Boxing

Many philosophers believe that one-boxing is irrational, and that the Principle of Expected Value Maximization should therefore be rejected. (Not all! One of my colleagues at MIT — who is also one of the smartest people I know — believes that when the predictor is 100% accurate, it is rational to one-box.)

In this section I'll tell you about some of the reasons that philosophers have set forth in defense of two-boxing.

Mathematosis

Let us begin with an analogy. Suppose there is a gene that has two different effects:

1. It increases the likelihood that one will do mathematics.
   (Perhaps the gene causes a certain hormone to be released, and the hormone can cause one to do mathematics.)
2. It increases the likelihood that one will suffer a terrible disease: 'mathematosis'.
   (The symptoms of mathematosis are too terrible to list here; let's just say you really, really don't want to have it.)

As a result of this, the disease is more prevalent amongst people who do mathematics than in the population at large. But this is **not** because doing matheamtics causes the disease (or because having the disease causes one to do mathematics). It is because doing mathematics and having the disease have a common cause: they are both caused by the gene.

(*Compare*: wet sidewalks are more likely at times when people are using umbrellas, but that's not because umbrella use causes wet sidewalks (or because wet sidewalks cause umbrella use), it is because wet sidewalks and umbrella use have a common cause: they are both caused by rain.)

Now suppose that you would quite enjoy doing mathematics. What should you do in a world with mathematosis? Should you refrain from doing mathematics even though you would enjoy it?

Of course not! If you carry the gene, you're likely to get the disease, but there is nothing you can do about it now. So better to do mathematics, and enjoy life while you are still healthy. And if you don't carry the gene, there is no need to worry: you won't get the disease, regardless of whether you do mathematics. So, again, there's no reason to refrain from enjoying yourself. Either way: you should do mathematics!

(*Compare*: Suppose you don't want the sidewalks to be wet. Should you refrain from using an umbrella? Of course not! If rain is on the way, the sidewalks will get wet regardless of whether you use your umbrella. So better to use it and stay dry. And if rain is not on its way, there is no need to worry: the sidewalks will remain dry regardless of whether or not you open your umbrella.)

Dependence, Causal and Probabilistic

Two events are **probabilistically independent** if the assumption that one of them occurs does not change the likelihood that the other one will occur. (Otherwise, each of them is **probabilistically dependent** on the other.)

Two events are **causally independent** if neither of them is a cause of the other. (Otherwise, the effect is **causally dependent** on the cause.)

The mathematosis example can be used to illustrate the difference between probabilistic dependence and causal dependence:

- Having the disease is probabilistically *dependent* on doing mathematics, because the assumption that you do mathematics increases the likelihood that you have the disease.
- Having the disease is causally *independent* from doing mathematics, because doing mathematics does not *cause* the disease. (What we have instead is a *common cause*: the gene causes both the disease and a desire to do mathematics.)

Something similar happens in the case of Newcomb's Problem:

- Whether or not the large box contains a million dollars is probabilistically *dependent* on your choice to one-box or two-box, because the assumption that you one-box increases the likelihood that the large box contains the money.
- Whether or not the large box contains a million dollars it is causally *independent* from your choice to one-box or two-box, because your action doesn't cause the box to have the money in it; the box is, after all, already sealed by the time you act. (Here too we have a *common cause*: your psychological constitution causes both your decision and the predictor's prediction.)

The reason that the Principle of Expected Value Maximization recommends one-boxing rather than two-boxing is that it uses probabilistic dependence, rather than causal dependence, to determine how much weight to give each of the possible outcomes of one's actions. (Notice, in particular, that in calculating the expected value of an action A one assigns weights to the possible outcomes of A by using the conditional probability $p(O_i|A)$, which tracks probabilistic dependence between O_i and A, rather than causal dependence.)

Philosophers who think that one-boxing is irrational think that that is precisely where the Principle goes wrong. Our decision making, they think, should be guided by causal dependence, not by probabilistic dependence. (I'll have more to say about this when we turn to Causal Decision Theory, below.)

Exercises

1. Alice and Bob make independent decisions to go for a walk in the park on Tuesday afternoon, and neither of them brings an umbrella. Let A be the event of Alice's getting soaked in rain, and B be the event of Bob's getting soaked in rain. Are A and B probabilistically independent? Are they causally independent?

- only probabilistically independent
- `only causally independent only causally independent`
  - The occurrence of A would make it much more likely that B occurs, since it raises the probability that it rained and therefore that they both got soaked. So A and B are not probabilistically independent.
  - Since A and B made their decisions independently, we have been given no reason to think that A's occurrence would cause B to occur, or that B's occurrence would cause A to occur. (If either of these events occurs, it will presumably be caused by the relevant subject's decision go for a walk without an umbrella, and by the presence of rain.) So we can expect A and B to be causally independent.
- both probabilistically independent and causally independent
- neither probabilistically independent nor causally independent

2. Let S be the event of Jones's smoking, and C be the event of Jones's being diagnosed with lung cancer. Are S and C probabilistically independent? Are they causally independent?

- only probabilistically independent

- only causally independent only causally independent
- both probabilistically independent and causally independent
- `neither probabilistically independent nor causally independent`
  - Smoking is a cause of lung cancer. So C is not causally independent on A.
  - Because smoking is a cause of lung cancer, Jones's smoking increases the probability that he will be diagnosed with lung cancer. So C and A are not probabilistically independent.

QUESTIONS

What If the Predictor is 100% Accurate?

Maybe We Should One-Box Because of Backwards Causation?

Temptation

Even if you are a committed two-boxer, it can be difficult to escape the temptation to one-box.

Imagine that a billionaire comes to town, and offers everyone the chance to participate in a Newcomb scenario. After the first day, your one-boxer friends are absolutely delighted. They have each found a million dollars in the large box, and they have spent the night celebrating with caviar and champagne.

Your two-boxer friends, on the other hand, are all crestfallen. They all found the large box empty, and although there's nothing wrong with getting a thousand dollars, it's not quite the same as a million.

You get your chance to participate the next day. When the time finally comes, you decide to two-box, and, predictably, you find nothing in the large box.

Your one-boxing friends can't stop laughing. "What a fool!" they cry. "What on Earth possessed you to two-box!?" Are they right? Did you make a mistake?

VIDEO REVIEW: Why Aren't You Rich Too?

My own view is that you did not make a mistake when you two-boxed.

The billionaire came to town with the intention of rewarding irrational behavior (i.e., one-boxing). So it would have certainly been in your interests to somehow make her believe that you are irrational. (And, of course, once the boxes were sealed, it would have been in your interests to take both boxes.)

But by the time the billionaire arrived in town she had already decided who would be rewarded and who would not. Your large box was empty from the start, and there's nothing you could have done to change that. Leaving the thousand dollars behind would certainly not have helped.

---

4. The Tickle Defense

---

5. Casual Decision Theory

---

6. The Prisoner's Dilemma

---

7. Further Resources

---

8. And One More Thing...