Newcomb’s problem

Today we begin our discussion of paradoxes of rationality. Often, we are interested in figuring out what it is rational to do, or to believe, in a certain sort of situation. Philosophers and others - including people in various social sciences working on rational decision theory - have tried to approach these sorts of questions of rationality systematically. This involves trying to formulate general rules which, when applied to a particular situation, will tell us the rational act to do or the rational belief to form.

The paradoxes of rationality are, typically, cases in which otherwise extremely plausible rules of this sort seem to inexplicably break down - or, in the case of paradoxes like the one we'll discuss today, in which two otherwise extremely plausible rules of rationality deliver contradictory answers.

Our topic today is **Newcomb's problem**. Newcomb's problem is named after William Newcomb, a physicist at the Livermore Laboratory in California - it's named after him because the philosopher Robert Nozick, who was the first to discuss the problem in print, credits the problem to him.

There are various different versions of Newcomb's problem; but an intuitive presentation of the problem is very easy to give.

There are various different versions of Newcomb's problem; but an intuitive presentation of the problem is very easy to give.

Suppose that you go to the St. Joseph's County fair, and you come across a wise looking man in a booth, who is offering fair-goers a chance at an unusual game. When you play his game, you are presented with two boxes - Box A and Box B. You have a choice about whether you will take the contents only of Box B, or the contents of **both** Box A and Box B.

The Predictor

Box A

Box B

You watch many, many fair-goers, many quite similar to yourself, play the game. And you notice two main things. First, the Predictor **always** places $10 in Box A.

Box B is a bit trickier. What you notice, after watching several thousand trials, is this: if the person playing the game chooses both boxes — if they "2 box" — then there is nothing in Box B. And then the person walks away with $10, since that is the sum of the two boxes.

But if the person chooses just Box B — if they "1 box" — then there is, invariably, $1000 in Box B. And so the people that 1 box — and, again, you have watched several thousand trials - always walk away with $1000.

What seems to be happening is this: the Predictor is very good at guessing whether players are going to 1 box or 2 box. If he guesses that they are going to 2 box, he puts nothing in Box B. If he guesses that they are going to 1 box, he puts $1000 in Box B.

You might think that there's some funny business with the boxes — that the Predictor or one of his cohorts puts money in or takes money out after the choice has been made. But you are able, through careful observation, to be absolutely sure that the box is closed, so that no money can enter or leave the box between the making of the choice and the opening of the box.

The Predictor

Box A

Box B

Now it is your turn. You walk up to the boxes. The Predictor looks at you knowingly. You think to yourself: “Whatever is in the boxes is already there; I might as well take both. I could use the extra $10, whatever ends up being in Box B.” But then you think again: “Every 1 boxer I have seen walks away with $1000, and every 2 boxer walks away with $10. I would be an **idiot** to choose both boxes!” What should you do?

The intuitive conflict here comes from a conflict between two different ways of making decisions under conditions of uncertainty.

One of these involves the rule of **expected utility.**

One of these involves the rule of **expected utility.**

It's useful to think about this rule in terms of a simple bet.

> I'm about to flip a coin, and offer you the following bet: if the coin comes up heads, then I will give you $5; if it comes up tails, you will owe me $3. You know that it is a fair coin. Should you take the bet?

We might represent this decision using the following table:

Courses of action	Possibility 1: Coin comes up heads	Possibility 2: Coin comes up tails
Take the bet	win $5	lose $3
Don't take the bet	$0	$0

There is a ½ probability that the coin will come up heads, and a ½ probability that it will come up tails. In the first case I win $5, and in the second case I lose $3. So, in the long run, I'll win $5 about half the time, and lose $3 about half the time. So, in the long run, I should expect the amount that I win per coin flip to be the average of these two amounts — a win of $1.

We can express this by saying that the **expected utility** of taking the bet is $1.

To calculate the expected utility of an action, we assign each outcome of the action a certain **probability**, thought of as a number between 0 and 1, and a certain **value** (in the above case, the relevant value is just the money won). In the case of each possible outcome, **we then multiply its probability by its value; the expected utility of the action will then be the sum of these results**.

Let's see how this works in the case of the simple bet just described.

To calculate the expected utility of an action, we assign each outcome of the action a certain **probability**, thought of as a number between 0 and 1, and a certain **value** (in the above case, the relevant value is just the money won). In the case of each possible outcome, **we then multiply its probability by its value; the expected utility of the action will then be the sum of these results.**

Let's see how this works in the case of the simple bet just described.

Courses of action	Possibility 1: Coin comes up heads	Possibility 2: Coin comes up tails
Take the bet	win $5	lose $3
Don't take the bet	$0	$0

We first have to ask: what are the probabilities of Possibility 1 and Possibility 2?

Since we are assuming that this is a fair coin, we can assume that the probability of each possibility is 0.5.

We then calculate the expected utility of taking the bet by multiplying the value of each outcome with the probability of that outcome, and summing the results, as follows:

Expected utility of taking the bet: ½ * $5 + ½ * (-$3) = $1

This by itself does not tell us whether to take the bet; we have to ask whether the expected utility of this course of action is greater than the expected utility of the other available course of action — in this case, not taking the bet. We calculate the expected utility of not taking the bet in the parallel way:

Expected utility of not taking the bet: ½ * $0 + ½ * $0 = $0

Since the expected utility of taking the bet > the expected utility of not taking it, it seems rational to take the bet in this case (supposing, of course, that one desires money, that there are no other costs of taking the bet, etc.).

Since the expected utility of taking the bet > the expected utility of not taking it, it seems rational to take the bet in this case (supposing, of course, that one desires money, that there are no other costs of taking the bet, etc.).

Generalizing from this sort of example, one might think that the following principle about rational action is quite plausible:

The rule of expected utility

It is always rational to pursue the course of action with the highest expected utility.

It is worth emphasizing that this principle does **not** say that you should always act so as to maximize the amount of money you have. It may be that in many cases you value certain other goods more than money, and that you would miss out on those goods by pursuing monetary gain. In this sense, the rule of expected utility is meant to be independent of any particular assignment of values to outcomes — however one values outcomes, the idea is that one can use the rule of expected utility to tell you what it is rational to do **given your choice of values**.

That said, it is often useful to simplify by focusing on cases in which the only relevant value is monetary gain, since one can simply plug the relevant dollar amounts into the calculations rather than trying to assign numeric values to the various outcomes. (This does ignore the fact that you might, for example, care more about the difference between $10 and $20 than about the difference between $1010 and $1000 — ignore that for now. We'll be discussing this fact when we discuss the St. Petersburg paradox.)

It's no great virtue of the rule of expected utility that it tells us what to do in the case of the simple coin bet described above — it was, after all, pretty obvious. But it can also tell us what to do in slight more complex cases, like this one:

You're playing a game of poker, and are trying to decide whether to match a bet of $10. If you match the bet, you will get to draw 1 more card; and, if you do so, you estimate that you have a 25% chance of drawing a straight. In games of poker with 4 players, a straight will win the pot half of the time. If you don't draw the straight, you have a losing hand. There is $88 in the pot.

What's the expected utility of matching the bet?

The rule of expected utility

It is always rational to pursue the course of action with the highest expected utility.

Now let's try to apply this rule to Newcomb's problem. How should we calculate the expected utility of 1 boxing and 2 boxing?

The first thing we need to do is figure out the probabilities of the two possible amounts - $0 and $1000 being in Box B. And a very natural thought is that, on the basis of our extensive experience at the fair, we should assign either the following probabilities, or something quite close to them:

The probability of $1000 in Box B, if I 1 box: 100%
The probability of $1000 in Box B, if I 2 box: 0%

The probability of $0 in Box B, if I 1 box: 0%
The probability of $0 in Box B, if I 2 box: 100%

This suggests the following expected utility calculation:

Courses of action	Possibility 1: $1000 in Box B + $10 in Box A	Possibility 2: $0 in Box B + $10 in Box A	Expected utility
1 Box	$1000	$0	100% * $1000 + 0% * $0 = $1000
2 Box	$1010	$10	0% * $1010 + 100% * $10 = $10

This chart is interestingly different from the other expected utility charts we have discussed, since here the probability of the outcomes differs depending upon the course of action chosen. But, setting this aside for now, the result seems completely decisive: the expected utility of 1 boxing is $1000, and the expected utility of 2 boxing is $10. Small changes in the probabilities — as if, for example, one thinks that the past experience at the fair should not make one **completely** certain, but only 97% sure, that all 1 boxers will get $1000 in Box B — would obviously not affect the overall result very much.

There thus appears to be a very strong expected utility argument for 1 boxing.

There is thus a very strong expected utility argument for 1 boxing.

But there is, it seems, an equally strong argument for the opposite conclusion, which uses an intuitively equally plausible principle about rational decision making as the rule of expected utility.

As above, we can get clearer on this principle by considering a simple bet:

> I offer you the chance of choosing heads or tails on a fair coin flip, with the following payoffs: if you choose heads, and the coin comes up heads, you win $5; if you choose heads, and the coin comes up tails, you lose $1. If you choose tails, then if the coin comes up heads, you get $2, and if it comes up tails, you lose $1.

Even if one knew nothing about expected utility, there would be a powerful argument in favor of choosing heads, as becomes clear if we think about the following chart:

Courses of action	Possibility 1: The coin comes up heads	Possibility 2: The coin comes up tails
Choose heads	win $5	lose $1
Choose tails	win $2	lose $1

One way to put the reason behind choosing heads is as follows: **there is one possibility on which you are better off having chosen heads, and no possibility on which you are worse off choosing heads.** This is to say that choosing heads **dominates** choosing tails.

In general, one choice A dominates another choice B if and only if under every possible condition, A leaves you no worse off than B, and in at least one condition, A leaves you better off than B.

The rule of dominance

If you are choosing between A and B, and A dominates B, you should choose A.

In general, one choice A dominates another choice B if and only if under every possible condition, A leaves you no worse off than B, and in at least one condition, A leaves you better off than B.

The rule of dominance

If you are choosing between A and B, and A dominates B, you should choose A.

But the rule of dominance, unlike the rule of expected utility, seems to point in favor of 2 boxing. For consider the following chart:

Courses of action	Possibility 1: The Predictor has placed $1000 in Box B (and $10 in Box A)	Possibility 2: The Predictor has placed $0 in Box B (and $10 in Box A)
1 box	$1000	$0
2 box	$1010	$10

As this chart illustrates, 2 boxing dominates 1 boxing. Hence it seems that, insofar as the rule of dominance seems quite plausible, there is a very strong argument in favor of 2 boxing.

So which rule should we follow? One thought, which Sainsbury pursues, is that we can make some progress by thinking more about how the case is supposed to work. In particular, we should ask: how does the Predictor always manage to get things right?

So which rule should we follow? One thought, which Sainsbury pursues, is that we can make some progress by thinking more about how the case is supposed to work. In particular, we should ask: how does the Predictor always manage to get things right?

One possibility is that the Predictor always manages to give the right answer because, after you decide whether to 1 box or 2 box, he causes the appropriate amount of money to have been placed in Box B before your choice. The idea is not that the Predictor, through sleight of hand, puts some money in the box after your selection - you managed to rule out the possibility of this - but that he, after your decision, effects a change in how things were before your decision.

For this to be possible, **backward causation** - causal relations in which the effect precedes the cause - must be possible. This is controversial. But assume that it is possible. Then would it be rational to 1 box or 2 box?

1 boxing would clearly be the way to go. But then why does the rule of dominance lead us astray in this case?

Recall the chart we used to illustrate the dominance reasoning in favor of 2 boxing:

Courses of action	Possibility 1: The Predictor has placed $1000 in Box B (and $10 in Box A)	Possibility 2: The Predictor has placed $0 in Box B (and $10 in Box A)
1 box	$1000	~~$0~~
2 box	~~$1010~~	$10

The problem seems to be that, in a clear sense, which possibility turns out to be actual is **not** independent of the course of action chosen. This is because, in the ‘backwards causation’ version of the Newcomb problem, 1 boxing **causes** Possibility 1 to be actual, and 2 boxing **causes** possibility 2 to be actual. This means that the bottom left and top right squares in our chart of outcomes do not describe real possibilities.

Courses of action	Possibility 1: The Predictor has placed $1000 in Box B (and $10 in Box A)	Possibility 2: The Predictor has placed $0 in Box B (and $10 in Box A)
1 box	$1000	~~$0~~
2 box	~~$1010~~	$10

The problem seems to be that, in a clear sense, which possibility turns out to be actual is **not** independent of the course of action chosen. This is because, in the ‘backwards causation’ version of the Newcomb problem, 1 boxing **causes** Possibility 1 to be actual, and 2 boxing **causes** possibility 2 to be actual. This means that the bottom left and top right squares in our chart of outcomes do not describe real possibilities.

But note that we would get the same result if we changed the case so that the Predictor backward-caused the appropriate amounts to have been placed in Box B only 95% of the time. Still, in this case, 1 boxing would be clearly correct - and again, because which possibility turns out to be actual is not causally independent of the course of action undertaken.

This seems to suggest a certain restriction on dominance reasoning. Perhaps we should only follow the rule of dominance when the probabilities of the relevant possibilities are causally independent of the choice made. That is, perhaps we should adopt the following rule:

The restricted rule of dominance

If you are choosing between A and B, and A dominates B, and the relevant possibilities are causally independent of the choice made, you should choose A.

This seems to suggest a certain restriction on dominance reasoning. Perhaps we should only follow the rule of dominance when the probabilities of the relevant possibilities are causally independent of the choice made. That is, perhaps we should adopt the following rule:

The restricted rule of dominance

If you are choosing between A and B, and A dominates B, and the relevant possibilities are causally independent of the choice made, you should choose A.

This is a weaker rule of decision making, in the sense that it applies to fewer cases. Now, even our original rule of dominance did not apply to every decision; there are many decisions (most interesting ones) in which no course of action dominates the others. But our restricted rule of dominance restricts the scope of the rule still further.

However, one case in which the restricted rule of dominance still seems to have application is a version of Newcomb's problem **in which we stipulate that there is no backwards causation going on**. In this case, 2 boxing dominates 1 boxing, and the relevant outcomes are causally independent of the choice made.

Then we still seem to have a conflict with the rule of expected utility. After all, as we saw, expected utility calculations seem to dictate that it is rational to 1 box:

Courses of action	Possibility 1: $1000 in Box B + $10 in Box A	Possibility 2: $0 in Box B + $10 in Box A	Expected utility
1 Box	$1000	$0	100% * $1000 + 0% * $0 = $1000
2 Box	$1010	$10	0% * $1010 + 100% * $10 = $10

And as above we would get much the same result if we let the Predictor be right only 95% of the time, rather than every time. Does this show that the rule of expected utility should be rejected, or at least restricted in some way so that it does not give us this result?

Courses of action	Possibility 1: $1000 in Box B + $10 in Box A	Possibility 2: $0 in Box B + $10 in Box A	Expected utility
1 Box	$1000	$0	100% * $1000 + 0% * $0 = $1000
2 Box	$1010	$10	0% * $1010 + 100% * $10 = $10

And as above we would get much the same result if we let the Predictor be right only 95% of the time, rather than every time. Does this show that the rule of expected utility should be rejected, or at least restricted in some way so that it does not give us this result?

The odd thing about the above way of calculating expected utilities is something mentioned above in passing: it permits the probabilities of the various outcomes which are used to calculate the utility of the two courses of action to differ depending on the course of action taken - **even if we assume that which possibility is actual is causally independent of the action undertaken**.

But perhaps we should not permit calculations of expected utility to work in this way. Perhaps we should adopt the following rule governing probability assignments to outcomes:

> **If the probability of an outcome O is causally independent of the choice between Act 1 and Act 2, then calculations of the expected utility of Act 1 and Act 2 must treat the probability of O as fixed.**

If we adopt this rule, then **the probabilities assigned to Possibilities 1 and 2 will be the same for the two courses of action undertaken**. And this means that the expected utility calculations will favor 2 boxing - in agreement with the restricted rule of dominance.

Should we adopt a rule of this sort?

But perhaps we should not permit calculations of expected utility to work in this way. Perhaps we should adopt the following rule governing probability assignments to outcomes:

> **If the probability of an outcome O is causally independent of the choice between Act 1 and Act 2, then calculations of the expected utility of Act 1 and Act 2 must treat the probability of O as fixed.**

If we adopt this rule, then **the probabilities assigned to Possibilities 1 and 2 will be the same for the two courses of action undertaken**. And this means that the expected utility calculations will favor 2 boxing - in agreement with the restricted rule of dominance.

Should we adopt a rule of this sort?

Certain real-world analogues of Newcomb problems suggest that we should. Suppose, for example, that I am wondering whether I have a certain disfiguring disease. I may know that people with this disease almost always lose weight. I might then consider the following line of reasoning:

> "I know that the probability of my having the disease is much less if I gain weight than if I lose weight. Hence the expected utility of my gaining weight is much greater than the expected utility of my losing weight. So, I should begin eating as much as I possibly can, to bring about a situation in which I gain weight."

There's obviously something wrong with this way of reasoning. One suggestion is that what is wrong with it is that it treats the probability of my having the disease as varying depending on whether I begin to eat excessively, despite the fact that my having the disease is causally independent of my choice about how much to eat.

One might say that loss of weight is a symptom, rather than a cause, of the disease. Just so, the my choosing 1 box would be a symptom, rather than a cause, of the Predictor's having chosen to put $1000 in Box B. On this view, 1 boxing is just as irrational as eating to decrease the odds of your having the disease.

We started out with the idea that Newcomb's problem is a case which exhibits a conflict between two eminently plausible principles of rational choice: the rule of expected utility, and the rule of dominance. What we've been arguing is that both of these principles need to be restricted in a certain way. Once they are restricted, they don't disagree in these cases. They agree that 1-boxing is the way to go in the backwards causation version of the case, and that 2-boxing is the way to go in the version of the case which involves no backwards causation.

There are two ways, however, to put pressure on the view that 2-boxing is the way to go in the central version of Newcomb's problem, which involves no backwards causation.

First, if we think about the 1 boxers and the 2 boxers, the 1 boxers invariably walk away richer. Given that the aim of playing this game with the Predictor is presumably to maximize your money, doesn't this make the 1 boxers right?

In response to this question, the 2 boxer might well concede that, given ample time, the rational course of action is to convince oneself to be a 1 boxer since, given the Predictor's acumen, that is the best way to get the Predictor to put $1000 in Box B. But that is consistent with the idea that, when one is actually faced with a choice between 1 boxing and 2 boxing — and one is certain that there is no backwards causation — the rational response is to 2 box.

A second way to call into the question the rationality of 2 boxing is to imagine the point of view of someone betting on the outcomes of games played with the Predictor. Wouldn't they be rational to bet that if you 1 box you will be better off than if you 2 box? And if it would be rational for them to bet this, why wouldn't it be rational for you — who have the same evidence — to believe it? But, if you do believe that you would be better off 1 boxing than 2 boxing, how can it be rational for you to 2 box?

As before, it is plausible that this challenge can be defused by noting that the 2 boxer can agree that it would be rational to, if possible, become the sort of person who would cause the Predictor to put $1000 in Box B — namely, a 1 boxer. But that does not change the fact that — according to this view — once the money is in the boxes, and the Predictor asks for your choice, the right answer is to 2 box.