The Problem

Two boxes are before you. One is transparent; you can see that inside is a thousand dollars in cash. The other box is opaque; you don’t know what’s in there. But you *do* know that it is either a *million* dollars in cash, or nothing.

You have two choices. You can take the opaque box only, and hence get either nothing or a cool million. Or you can take *both* boxes, and hence get either a thousand dollars or one million, one thousand dollars.

Seems like an easy choice, right? Take both! There seems to be *no reason at all* to take just one.

Here is the twist: you have some information about how it was decided whether or not to put a million dollars in the opaque box. 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 opaque box), or a **two-boxer** (i.e. the kind of person who would take both boxes). If the expert concluded that you are a one-boxer, the opaque box was filled with a million dollars. If she concluded that you are a two-boxer, the opaque box was left empty.

In other words:

Verdict of Expert Contents of Large Box Contents of Small Box You are a One-Boxer $1,000,000 $1,000 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!)

For Two-Boxing: Dominance

Dominance

What should you do, when you are faced with a situation with uncertain outcomes?

Suppose you need to decide how to get to New York. Both are equally convenient, given your final destination, and they are about the same price. Your main concern is to minimise time.

How much time you will take depends on things you don’t know. There is a blizzard brewing; it may or may not strike while you are in transit. Here is a matrix representing the **outcome** for each possibility.

plane-no-blizzard 1 plane-blizzard 2 train-no-blizzard 4 train-blizzard 6

The cells at the left hand side of the matrix represent the possible **acts** among which you are choosing. The cells at the top of the matrix represent what are called the **states**. How good an act is depends on which state the world is in.

(In general: specifying a decision problem involves specifying the acts, the states, and the outcome for each act-state pair. We will, in general, specify a decision problem using a table like this: State 1 State 2 etc. Act 1 Outcome 1–1, Outcome 1–2, etc. Act 2 etc. This table is called a **decision matrix**)

Let us assign a number to each outcome, representing how much you want each outcome to occur:

plane-no-blizzard 5 plane-blizzard 4 train-no-blizzard 2 train-blizzard 0

These numbers are called **values** (or sometimes “utilities”). The exact number you give to each outcomes won’t matter, for any purpose we talk about here; what is important is the *ratios* among the values in a given decision problem.

What should you do? Take the plane, obviously! It is the better choice if there’s a blizzard, and it’s the better choice if there *isn’t* a blizzard. Either way, it’s the better choice. So it’s the better choice.

When an act A is better then an act B whatever the state, act A is said to **dominate** choice B.

The **Dominance Principle** is the following rule for how to make decisions under uncertainty:

> If an act A dominates all other options, choose A!

Seems like a pretty good principle, right?

Dominance and the Newcomb Problem

The main argument for two-boxing is a dominance argument.

Consider the decision matrix for Newcomb's Problem.

Assuming that you want a thousand dollars more than nothing, and $1,001,000 more than $1000, then two-boxing *dominates* one-boxing. So, by the Dominance Principle, you should two-box!

For One-Boxing: Expected Value

What should you do, when you are faced with a situation with uncertain outcomes?

Consider again your train/plane decision above. Let's suppose you can make some estimate about how likely this blizzard is; .3 chance of a blizzard, let's say, so a .7 chance of no blizzard.

With those numbers, we can calculate the *expected value* of taking the plane, and the expected value of taking the train. The expected value of taking the plane is the weighted average of the value if there's a blizzard, and the value if there isn't one. The weights are the probability of blizzard and no-blizzard, respectively. So the expected value of taking the plane is:

Similarly, the expected value of taking the train is:

What should you do? Take the plane, obviously! That's the outcome that you can expect to do best by taking.

In general, the **expected value (version 1)** of a choice is

(Why "version 1"? Stay tuned…)

The **Principle of Expected Value Maximisation (PEVM)** is the following rule for how to make decision under uncertainty:

> Always make the choice with the highest expected value!

Compare this principle to the Dominance Principle. It is more general; it tells you what to do in *every* decision under uncertainty, not just what to do when one choice dominates all the others. But it requires more inputs; you need to supply some probabilities as well as values, whereas the Dominance Principle requires only values.

Seems like a pretty good principle, right?

A Bad Argument

You have a heart condition. Without treatment, you are likely to die young. But there is a very safe surgical procedure that can significantly reduce the chance that you will die young. It isn’t perfect; in some cases, it makes no difference, and the patient dies young anyway. But usually, people with your condition who have the treatment go on to have long, healthy lives.

You are considering: should you have the surgery? Let’s draw up the decision matrix…

How do you value the outcomes here? Well, of course, you really would rather not have surgery than have it, all else equal — who likes having surgery? But you really, *really* would prefer to live a long, healthy life then to die young. So the way you value the outcomes, in this scenario, is something like this:

No surgery live–1000 No surgery die–50 Surgery live 950 surgery die–0

But notice this: in the above matrix, *No Surgery* dominates *Surgery*. So, by the dominance principle, you should not have the surgery. Does that sound right?

No it does not! And it doesn't sound right because it *isn't* right; the Dominance Principle is giving you bad advice, in this case.

Does the PEVM do any better? To use the PEVM, we first need to supply some probabilities; in particular, the probability that die young versus the probability that you live long and prosper. But if you think about it for a moment, you'll see there is something strange about coming up with those probabilities. How likely it is that you die young depends on how likely it is that you have the surgery. And you are trying to figure out whether to have the surgery or not right now; it seems strange to rate how likely it is that you will have surgery as part of your deliberation about whether or not to have surgery. It's not clear it even makes *sense* to do that.

We can change the way we calculate expected values to take care of this. We can calculate using the *conditional* probabilities of the states, on the condition that we act a certain way, instead of the absolute probabilities of the states. The formula for **expected value (version 2)** is this:

To use *this* formula we need to know the probability that you live long, given that you have the surgery, the probability that you die young, given that you have the surgery, the probability that you live long, given you don't have the surgery, and the probability that you die young given that you don't have the surgery. Let's make some estimates:

We then make the calculation…

And lo and behold, we get the sensible answer. So this round, PEVM wins.

The PEVM and Newcomb's Problem

Let us look again at the decision matrix for Newcomb's problem.

We have some information on the conditional probability of the states, given your acts; it follows from the accuracy of the predictor. The probability of there being a million in the opaque box, given that you two-box, is 1%; the probability if there being nothing in the opaque box, given that you two-box, is 99%. And so on.

So the expected value of Two boxing is …. and the expected value of one-boxing is … So, clearly, by the PEVM, you should one-box!

The Fallacy of the Bad Argument

Consider again the bad argument for not having surgery. The fan of the PEVM has a diagnosis of what went wrong: we were dealing with a situation in which the states were not *evidentially independent* of the acts. Dominance arguments are no good in such situations.

What does that mean? Let me explain.

In decision problems, when we are talking about “probability” of a state, we mean something like the confidence of the person making the decision that that state obtains. (We will talk more about what *exactly* we mean in Topic 6 of this course.) That the probability of a state is a certain number is a fact about the agent, and how things seem to her, given the information she has.

Similarly, the *conditional* probability of a state B on an act A is (roughly) the confidence the agent *would* have in B if she were to learn A.

When the conditional probability P(B|A) is equal to P(B), then B is said to be **evidentially independent** of A. That means: learning that A would have no effect on how confident the agent is that B.

The train versus plane example is a good example of evidential independence. This isn’t *necessarily* true, but in any realistic case, the agent will have no reason to change her confidence that a blizzard will occur if she learns that she will take the plane. Those two propositions are epistemically independent for the agent (and for basically any agent).

When the states are evidentially independent of the acts, version 1 and version 2 of the expected value calculation are the same. It is only when the states are *not* evidentially independent of the acts that they differ. Similarly, the only time that the Dominance Principle gives conflicting advice with the PEVM is when the states are not evidentially independent of the acts.

What the above example, with the surgery, seems to show is that in these cases, where the states are *not* evidential independent of the acts, the PEVM gets things right, and the Dominance Principle gets things wrong.

The Fallacy and Newcomb's Problem

For the one-boxer, the two-boxer is making the same mistake as the person who argues that you should avoid the surgery; they are using dominance reasoning in a case where the states are not evidentially independent of the acts. But just as that reasoning gives bad verdicts in the surgery case, it gives bad verdicts in Newcomb's problem.

For Two-Boxing: Evidential versus Causal Dependence

The two-boxer has a retort. You are indeed correct, says the two-boxer, about the surgery case; that is a bad application of the Dominance Principle. But you have misdiagnosed the problem. The problem is not that the states are *evidentially* dependent on the acts. The problem is that the states are *causally* dependent on the acts. Dominance reasoning does not work when the states are causally dependent on the acts, but it *does* work otherwise, even if in cases involving evidential dependence without causal independence.

Let me, the two-boxer, show you why.

Mathematosis

Suppose there is a gene that has two different effects:

1. It increases the likelihood that you will study mathematics.
2. It increases the likelihood that one will suffer a terrible disease: mathematosis.

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

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

This mathematosis case is a case where there is evidential dependence between states and acts but no causal dependence between states and acts.

Now, suppose you would like to study mathematics, but you really, really don’t want to have mathematosis (the symptoms are truly horrible, though they strike later in life). Should you study mathematics?

The decision matrix is something like this:

study-no-disease–1000 study-disease–10 no-study-no-disease–990 no-study-disease–0

And let us say that you have numbers of the prevalance of mathematosis; it is pretty common in people who study mathematics — say, 0.3 of people who study mathematics have it. But it is extremely rare in the population at large — maybe 0.0001.

The Dominance Principle says: study! The PEVM says: don’t study! Which is right?

The Dominance Principle, obviously! If you carry the gene, you’re relatively likely to get the disease, but there is nothing you can do about it now. So better to study 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 study mathematics or not. So, again, there’s no reason to refrain from enjoying yourself. Either way: you should study mathematics!

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

Dependence, Causal and Probabilistic

To summarise:

Having the disease is evidentially dependent on studying mathematics, because the assumption that you study 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 evidentially 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 evidential 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 you assigns weights to the possible outcomes of A by using the conditional probability p(Si|A), which tracks evidential dependence between Si and A, rather than causal dependence.)

The two-boxer thinks that that is precisely where the PEVM goes wrong. Our decision making, they think, should be guided by causal dependence, not by probabilistic dependence.

Causal Decision Theory

Decsion theory is a theory about what to do in decision problems. All versions of decision theory endorse the PEVM. But different versions use different ways of calculating expected value.

The first rigorous version of decision theory used version 1, above, of the expected value formula. Clearly, given cases like surgery, we want to do better than that, but how? One way is the one-boxer way, which uses version 2 of the expected value formula. The version of decision theory that uses that formula is called **evidential decision theory**, because the formula uses the conditional probability P(S|A), which tracks evidential dependence.

The two-boxer also has a way of calculating expected value. Rather than using conditional probabilities, it uses the probability of a certain conditional claim, called a “counterfactual conditional”. The kind of decision theory you get, if you calculate things the two-boxer way, is called **causal decision theory**. Let me explain.

Two Kinds of Conditionals

The easiest way of getting a handle on causal decision theory is to start by considering the difference between **indicative conditionals** and **counterfactual conditionals**. Here is an example of an indicative conditional, and a corresponding counterfactual conditional:

[Indicative Conditional] If you start working on Friday, you'll be done by Monday.

[Counterfactual Conditional] Had you started working on Friday, you would have been done by Monday.

These two conditionals have something important in common: they both claim that there is some sort of connection between your working on Sunday and your being done by Monday. There is, however, an important difference between them.

I can say the counterfactual conditional even if it is now Tuesday, and I know that, as a matter of fact, you didn't start working on Friday. Even if that is so, it might still be true that *had* you started working on Friday, you *would have been* done on Monday. In asserting the counterfactual conditional, I am considering a counterfactual alternative to the actual world — an alternative in which you *did* start working on Friday — and saying something about what that alternative world would have been like.

In contrast, it would be very strange for me to assert the indicative conditional on Tuesday, when I know that the student didn't start work on Friday. The indicative conditional is much more normal to assert when it is before Friday, and I don't know yet whether you will start working on Friday or not. In asserting the indicative conditional, I am considering the hypothesis that you will, in fact, in the actual world, start work on Friday, and saying something about what the world must actually be like, if that hypothesis is correct.

(In general, indicative sentences are expressed in English in the indicative mood, and counterfactual sentences in what passes for the subjunctive mood, in English. So words like “were”, “would” and “had” are pretty reliable indicators of a counterfactual, rather than indicative, conditional.)

Here’s another pair of conditionals, to help make the contrast clear:

1. If Oswald didn’t kill Kennedy, somebody else did.
2. If Oswald hadn’t killed Kennedy, somebody else would have.

Is sentence (1) true? Yes! *Somebody* shot Kennedy; if it wasn’t Oswald, it was someone else. That’s something we know about the actual world.

Is sentence (2) true? Assuming that Oswald did, in fact, kill Kennedy, and he was acting alone, then probably not. There is no particular reason to think that in an alternative scenario in which, contrary to fact, Oswald didn’t shoot Kennedy, then some other person would have done it.

Conditionals and the Newcomb Problem

There is often an important correlation between the truth of an indicative conditional and the truth of the corresponding counterfactual conditional.

Consider again our pair of conditionals:

[Indicative Conditional] If you start working on Friday, you'll be done by Monday.

[Counterfactual Conditional] Had you started working on Friday, you would have been done by Monday.

Suppose, first, that today is Thursday, and that you have a problem set due on Monday. Suppose, moreover, that the indicative conditional above is true: if you start working on Friday, you'll be done by Monday.

Now suppose that Friday comes and goes and that you don't, in fact, do any work. What does the truth of the indicative conditional on Thursday tell us about the truth of the subjunctive conditional on Monday morning? It is natural to think that it *guarantees* that the counterfactual conditional will be true on Monday. In other words: it is true on Monday that had you started working on Friday, you would have been done by Monday.

One of the things that make Newcomb scenarios weird, is that indicative and counterfactual conditionals about the Newcomb scenario don’t follow this pattern.

If you one-box, the predictor will almost certainly have predicted that you will one-box. So that the following indicative conditional will be true at a time before you make your choice:

[Indicative Conditional] If you one-box, you’ll almost certainly find a million dollars in the large box.

Now suppose that you nonetheless decide to two box, and that — as expected — you find the large box empty. Will the corresponding counterfactual conditional be true after you’ve made your choice?

[Counterfactual Conditional] Had you one-boxed, you would have almost certainly found a million dollars in the large box.

No! We now know that the large box was empty from the start, and your decision to one-box wouldn’t have changed that.

In general, indicative conditionals track evidential relationships, and are closely related to conditional probabilities;[1] whereas counterfactual conditionals track causal relationships. Usually, if A causes B, then “If A were to happen, B would happen” is true, and if A does not cause B, then “If A were to happen, B would happen” is false.[2]

The indicative and the counterfactual conditionals come apart, in the

Newcomb problem, because the evidential and the causal relationships come apart.

Causal Expected Value

According to the two-boxer, you should calculate the expected value of an act using the **expected value formula, version 3**

The arrow thing there represents the counterfactual conditional “If A were to happen, B would happen”.

It is straightforward to define conditional probabilities, like the ones used in version 2 of the expected value formula, in terms of the probabilities of A and B. It is *not* straightforward to define the probability of a counterfactual conditional in terms of the probabilities of A and B. It is, in fact, quite complicated to say exactly how you should think of the probability of A->B. We won’t get into the details here. Just rest assured that there is an alternative way of thinking about expected utility: the causal way.

For One-Boxing: the Tickle Defence

For One-Boxing: Why

Ain’t You Rich?

For Two-Boxing: Being Right Doesn’t Always Pay

The Deepest Issue

[1] Some philosophers think that indicative conditionals are, in a sense, expressions of conditional probabilities, in fact. To learn more about this, and about indicative conditionals in general, you should read this article by Dorothy Edgington, who is perhaps the leading authority on indicative conditionals. ↩

[2] Usually, but not always. The relationship between counterfactual conditionals and causation is actually quite complicated; but the above is true to a first approximation. ↩