Paradoxes
Lecture Six

Newcomb's Paradox

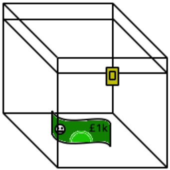
Rob Trueman
rob.trueman@york.ac.uk

University of York

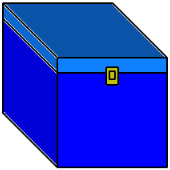
Newcomb's Paradox

One Box or Two?

Box A



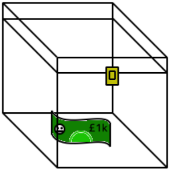
Box B



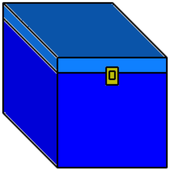
- You are presented with two boxes

One Box or Two?

Box A



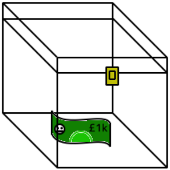
Box B



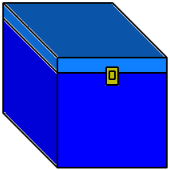
- Box A is transparent, and you can see that it contains £1,000

One Box or Two?

Box A



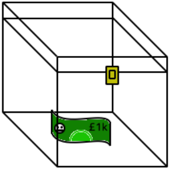
Box B



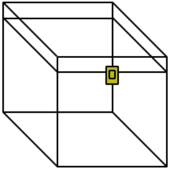
- Box B is opaque, and you cannot see what is in it

One Box or Two?

Box A



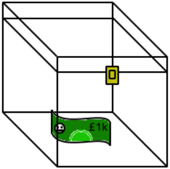
Box B



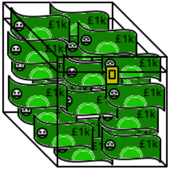
- You know that Box B is either empty...

One Box or Two?

Box A



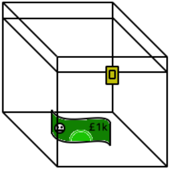
Box B



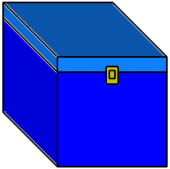
- ...or it contains £1,000,000...

One Box or Two?

Box A



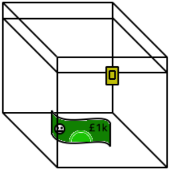
Box B



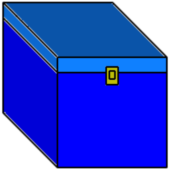
- ...but you do not know which

One Box or Two?

Box A



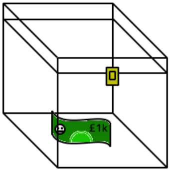
Box B



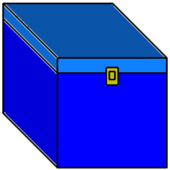
- You are made an offer:

One Box or Two?

Box A



Box B



- You may either take Box B, or take **both** Box A **and** Box B

The Predictor

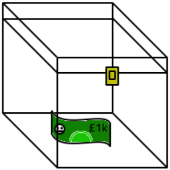
- One week ago, a woman known as the Predictor made a prediction about whether you would take one box or two boxes
- If she predicted that you would only take Box B, she put the £1,000,000 in B
- But if she predicted that you would take both Boxes A and B, she put nothing in B

The Predictor

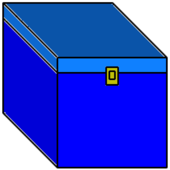
- The Predictor based her prediction on information about you that was available to her last week
- She didn't do it by magically looking into the future or anything like that
- She just found out everything she could about you to learn about your character, and based her prediction on that
- However, the Predictor is a **very** good judge of character
- She has played this game with lots and lots of people, and her predictions have always been right

One Box or Two?

Box A



Box B



- So now: will you take both boxes, or just Box B?

An Argument for Two-Boxing

- Either the Predictor put £1,000,000 in Box B, or she didn't
- Nothing you can do now will change which of these things the Predictor did a week ago
- If the Predictor put £1,000,000 in Box B, then you end up with more money if you take both boxes
  - You get £1,001,000 rather than £1,000,000
- If the Predictor put nothing in Box B, then you end up with more money if you take both boxes
  - You get £1,000 rather than nothing

An Argument for Two-Boxing

- So whatever the Predictor predicted, you are better off if you take both boxes
- So you should take both boxes!

An Argument for One-Boxing

- The Predictor has a **perfect** track record
- Lots of people have played her game, and you have seen all the results
- All the one-boxers are millionaires now, and all the-boxers came away disappointed
- The chances of you coming away a millionaire if you take just one box are very high
- The chances of you becoming a millionaire if you take both boxes are very low
- So you should take just Box B

Newcomb's Paradox

- So we have two arguments
  - One which tells us that we should take both boxes
  - And one which tells us that we should only take Box B
- When we put these two arguments together, we reach a contradictory conclusion:
  - You should take both boxes, but also you should only take one
- This is known as **Newcomb's Paradox**

Newcomb's Paradox

The Argument for One-Boxing Again

- We can represent all the possible combinations of actions and outcomes in the Predictor's game with the following table:

	Box B is empty	Box B is not empty
Take one box	£0	£1,000,000
Take both boxes	£1,000	£1,001,000

- Whether or not Box B is empty, you are better off (by £1,000) if you take both boxes
- So you should take both boxes!

The Dominance Principle

- This argument tacitly relies upon the **Dominance Principle**:

  DP If one course of action is no worse than its alternatives in any of the possible outcomes, and is better than its alternatives in at least one possible outcome, then it is rational to adopt that course of action

- Two-boxing "dominates" one-boxing

	Box B is empty	Box B is not empty
Take one box	£0	£1,000,000
Take both boxes	£1,000	£1,001,000

- In all of the possible outcomes, taking both boxes is **better** than taking one box

Utilities

- Situations can be given a **utility** score
- This score is a measure of the "goodness" of that situation
- We can base these scores on any criteria of "goodness" that you like
- However, it is often helpful when we can score the "goodness" in financial terms, because we then have some concrete numbers to work with
  - Consider the situation where you open just one box, and walk away with £1,000,000
  - We can say that that situation has a **utility** of 1,000,000

Maximising Expected Utility

- The expected utility of **outcome** O, relative to action A, is the utility of O given A, multiplied by the probability that O will occur, given that action A is performed:
  - $U_A(O) \times Pr(O|A)$
- The expected utility of an **action** is the sum of the expected utility of each outcome relative to that action:
  - $EU(A) = (U_A(O_1) \times Pr(O_1|A)) + (U_A(O_2) \times Pr(O_2|A)) + (U_A(O_3) \times Pr(O_3|A)) + ...$
- According to the principle of **Maximising Expected Utility** (MEU), it is rational to act so as to maximise expected utility
- It is this MEU which tells us to take just one box

Maximising Expected Utility

- Let's suppose that the Predictor has successfully played her game lots of times
- You are very confident that if you take just one box, she will put £1,000,000 in Box B, and if you take both boxes she will leave Box B empty
- To make things precise, let's pretend that the situation is like this

  $Pr(O_1|B) = 0.1; Pr(O_2|B) = 0.9$
  $Pr(O_1|A\&B) = 0.9; Pr(O_2|A\&B) = 0.1$

  - O_1 = Box B is empty
  - O_2 = Box B contains £1,000,000
  - B = You take Box B only
  - $A\&B$ = You take Box A and Box B

Maximising Expected Utility

$$Pr(O_1|B) = 0.1; Pr(O_2|B) = 0.9$$
$$Pr(O_1|A\&B) = 0.9; Pr(O_2|A\&B) = 0.1$$

- In the case of action B:
  - $U_B(O_1) =$£0; $U_B(O_2) =$£$1,000,000$
  - $EU(B) = (U_B(O_1) \times Pr(O_1|B)) + (U_B(O_2) \times Pr(O_2|B)) =$ $(£0 \times 0.1) + (£1,000,000 \times 0.9) = £900,000$

Maximising Expected Utility

$$Pr(O_1|B) = 0.1; Pr(O_2|B) = 0.9$$
$$Pr(O_1|A\&B) = 0.9; Pr(O_2|A\&B) = 0.1$$

- In the case of action $A\&B$:
  - $U_{A\&B}(O_1) =$£$1,000$; $U_{A\&B}(O_2) =$£$1,001,000$
  - $EU(A\&B) = (U_{A\&B}(O_1) \times Pr(O_1|A\&B)) + (U_{A\&B}(O_2) \times Pr(O_2|A\&B)) =$ (£$1,000 \times 0.9) + ($£$1,001,000 \times 0.1) =$ £$101,000$

Maximising Expected Utility

- The expected utility of taking just Box B is £900,000
- The expected utility of taking both Boxes A and B is £101,000
- So MEU tells you to take just Box B!

When Two Principles Go To War

- Newcomb's Paradox plays these two rules of reason against each other:

  DP If one course of action is no worse than its alternatives in any of the possible outcomes, and is better than its alternatives in at least one possible outcome, then it is rational to adopt that course of action

  MEU It is rational to act so as to maximise expected utility

- These both look like good rules, but they lead to different conclusions about what it is rational to do in the Predictor's game

Newcomb's Paradox as Premise-Flawed

- It seems, then, that Newcomb's Paradox is a demonstration that at least one of these principles is false
  - Or at least, false when taken as a completely general truth!
- Thus Newcomb's Paradox is premise-flawed: it relies on at least one false premise
- But which premise is false? Is it DP, MEU, or both?

Newcomb's Paradox

Never Study for an Exam!

- Here is a proof that it is **never** rational to waste your time studying for an exam:

	You pass the exam	You fail the exam
You study	2	0
You do not study	3	1

- Not studying dominates studying: whether your pass or fail the exam, you are happier if you did not study
- So by DP, it is rational for you not to study

Ignore that last Proof!!!

- You should of course study for your exams
- Whether or not you study affects **how likely** you are to pass or fail
- **DP totally ignores this fact!**
- So what the "proof" that you should not study for an exam *really* shows is that DP fails when our actions **causally affect** the probabilities of the possible outcomes

When DP Fails

- When my choice of action **causally affects** the probability of an outcome, then DP may break down, and MEU may become preferable
- On the other hand, when the probability of an outcome remains the same whatever action I perform, then MEU itself tells us to pick the dominant action:
  - $Pr(O_i|A_j) = Pr(O_i)$, for each outcome O_i and action A_j
  - So $EU(A_j) = (U_{A_j}(O_1) \times Pr(O_1)) + (U_{A_j}(O_2) \times Pr(O_2)) + ...$
  - The action with the highest expected utility in this situation will just be the dominant action

MEU Wins?

- At this point it might seem like the solution to Newcomb's Paradox is just to accept MEU, and reject DP
- DP only works when our actions do not causally affect the probabilities of the possible outcomes
- But when our actions do not affect the probabilities of the possible outcomes, MEU gives us the same result as DP
- But things are not so simple
- Sometimes our actions can in some sense "affect" the probabilities of possible outcomes without **causally** affecting them

Newcomb's Paradox

Should You Stop Smoking?

- We all know that there is a very strong statistical correlation between smoking and getting lung cancer
- We also all know that smoking **causes** lung cancer
- But imagine that things were really like this:
  - There is a gene which causes cancer in the vast majority of people who have it
  - This gene also causes the vast majority of people who have it to smoke
  - But smoking itself does not cause cancer

Should You Stop Smoking?

- In this situation, it is really bad news if you find yourself smoking
- If you find yourself smoking, then that should increase the probability that you have the gene which makes you smoke
- And that increases the probability that you are going to get lung cancer, because that gene also causes lung cancer
- But, I take it, you are not being irrational if you continue to smoke
  - Smoking doesn't **cause** cancer, the gene does, and even if you force yourself not to smoke, you will still have the gene
- So in this scenario, you may as well carry on smoking

Should You Stop Smoking?

- But MEU tells us, wrongly, that it **is** irrational for you to carry on smoking in this scenario

$$Pr(C|S) = 0.9$$
$$Pr(C|\neg S) = 0.2$$

  - C = You get lung cancer
  - S = You smoke

$$U(C) = -1,000,000$$
$$U(\neg C) = 0$$

$$EU(S) = (Pr(C|S) \times U(C)) + (Pr(\neg C|S) \times U(\neg C)) = -900,000$$

$$EU(\neg S) = (Pr(C|\neg S) \times U(C)) + (Pr(\neg C|\neg S) \times U(\neg C)) = -200,000$$

When MEU Fails

- In this scenario, the probability of your getting cancer is affected by your smoking, in the following sense:
  - If you discover yourself smoking, then you should increase the probability that you assign to your getting cancer
- But your smoking does not **causally** affect the probability of your getting cancer
- In general, MEU fails whenever your actions **affect** the probabilities of the possible outcomes, but do not **causally affect** them

Newcomb's Paradox

DP and MEU are both False

- Newcomb's Paradox played two principles against each other:
  - DP: If one course of action is no worse than its alternatives in any of the possible outcomes, and is better than its alternatives in at least one possible outcome, then it is rational to adopt that course of action
  - MEU: It is rational to act so as to maximise expected utility
- We've now seen that **both** principles are not true (when taken as universal rules)
  - DP doesn't work when our actions causally affect the probabilities of the possible outcomes
  - MEU doesn't work when our actions **do affect** the probabilities of the possible outcomes, but **do not causally affect** them
- So Newcomb's Paradox is premise-flawed twice over!

One Box or Two?

- But what **should** you do in the Predictor's game? Should you take one box or two?
- This is a bit like the smoking case we discussed a moment ago
- Whether we take one box or two **does** affect the probability that there is £1,000,000 in Box B, but it does not **causally** affect that probability
  - Remember, the Predictor made her prediction last week!
- So we should ignore MEU here
- Moreover, since there is no causal interaction in this case, we have given no reason not to trust DP
- So we should take DP's advice, and take both boxes

Newcomb's Paradox

Why Should We Care about Newcomb's Paradox?

- Isn't Newcomb's Paradox a bit silly?
- Nothing like the Predictor's game could ever happen, could it?
- According to Lewis (1979), Newcomb's Paradox is structurally identical to a very simple, real life problem, known as the **Prisoners' Dilemma**

The Prisoners' Dilemma

- You and a partner committed a bank robbery
- You are caught by the police, and taken to separate interrogation rooms
- The police have concrete evidence to convict you and your partner of some minor crimes, which will be enough to put you both away for 1 year
- But they don't have enough evidence to convict either of you for the bank robbery, which is a much more serious crime
- So the police make the following offer to you and your partner, and let you both know that they are offering it to both of you

The Prisoners' Dilemma

- If you confess, and your partner doesn't, then you'll go free as a reward, and your partner will go to prison for 10 years
- If you and your partner both confess, then you'll both go to prison for 7 years

	Partner confesses	Partner remains silent
You confess	7	0
You remain silent	10	1

The Prisoners' Dilemma

	Partner confesses	Partner remains silent
You confess	7	0
You remain silent	10	1

- Confessing dominates remaining silent: whatever your partner does, you are better off if you confess
- Moreover, nothing you can do will **causally** affect what your partner does, because you've been separated
- So this looks like exactly the sort of situation when DP should be obeyed, and you should confess

What does this have to do with Newcomb's Paradox?

- Although what you do does not **causally** affect what your partner does, it **does** affect the probability of what your partner does
- You and your partner are (we can imagine) both rational and similar in various ways, and so are very likely to act in the same way
- Thus, if you remain silent, the odds that your partner will also remain silent are increased
- If the probability that your partner will remain silent given that you do is sufficiently high, MEU tells you that you should remain silent

You Should Confess, and Take Two Boxes

- This is just like Newcomb's Paradox
- We have a weird situation where your actions affect the probabilities of the possible outcomes, but in a totally non-causal way
- So if we were right to say that in Newcomb's Paradox, you should follow DP, ignore MEU, and take two boxes, then we must likewise say that in the Prisoners' Dilemma, you should confess

The Prisoners' Dilemma is a Newcomb Paradox

> *Prisoners' Dilemma is a Newcomb Problem — or rather, two Newcomb Problems side by side, one per prisoner. Only the inessential trappings are different.*
>
> *Lewis (1979) p. 251*

What About Collective Action?

- Or that is what Lewis thought, anyway
- We might think that there are important differences between Newcomb's Paradox and the Prisoners' Dilemma
- Although no one prisoner can do anything to get fewer than 7 years in prison, together they have the power to make sure that they only go away for 1 year
- Hurley (1991) thinks that this is an important difference, and we'll be talking about that in the seminar

For the Seminar

- Required reading:
  - Lewis, D (1979) 'Prisoners' Dilemma is a Newcomb Problem', *Philosophy & Public Affairs* 8: 235–40
  - Hurley, S (1991) 'Newcomb's Problem, Prisoners' Dilemma, and Collective Action', *Synthese* 86: 173–96
- Both of these articles are available via the Reading List on the VLE

Next Week

- Next week we will be looking at Russell's Paradox
- Required reading:
  - *Paradoxes* Chapter 6, §6.1