Rationality, Morality and Economics

Topic 3, Lecture 1

Newcomb's Problem

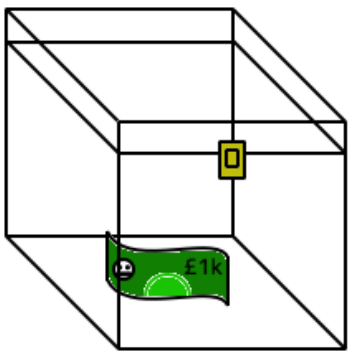
Rob Trueman
rob.trueman@york.ac.uk

University of York

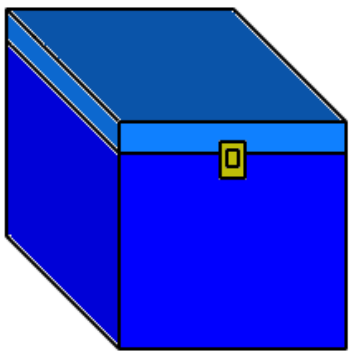
Newcomb's Problem

One Box or Two?

Box A



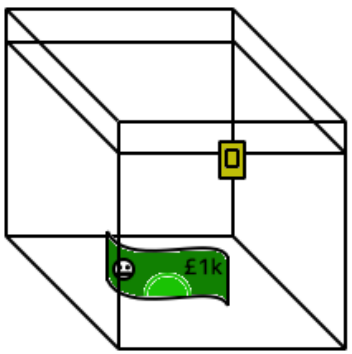
Box B



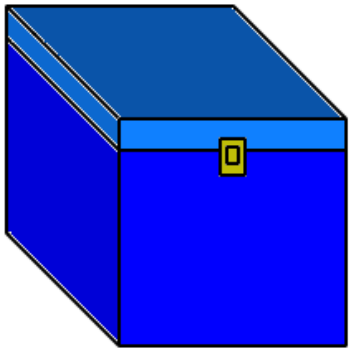
- You are presented with two boxes

One Box or Two?

Box A



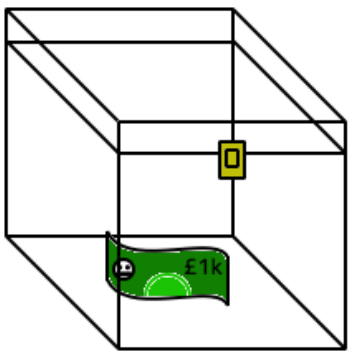
Box B



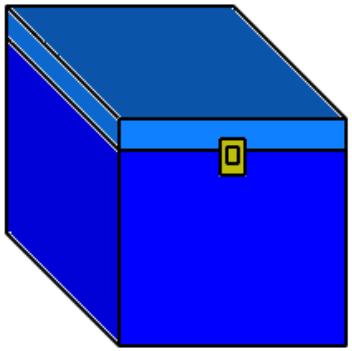
- Box A is transparent, and you can see that it contains £1,000

One Box or Two?

Box A



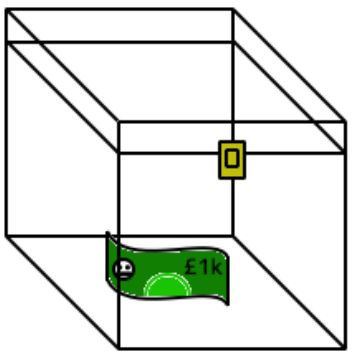
Box B



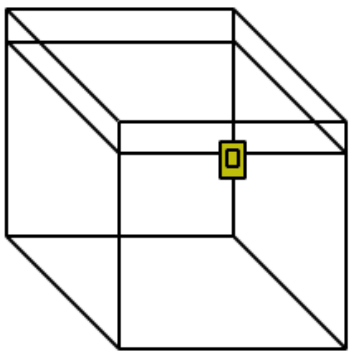
- Box B is opaque, and you cannot see what is in it

One Box or Two?

Box A



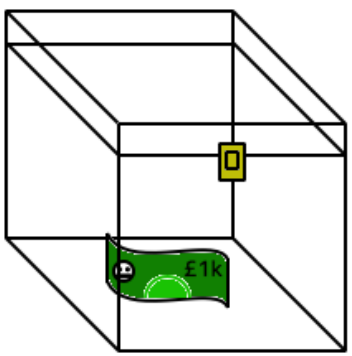
Box B



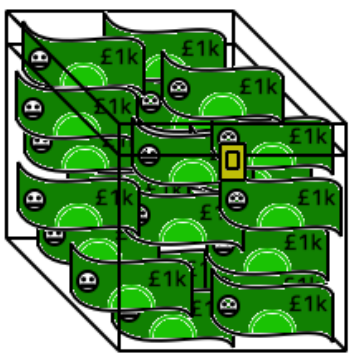
- You know that Box B is either empty...

One Box or Two?

Box A



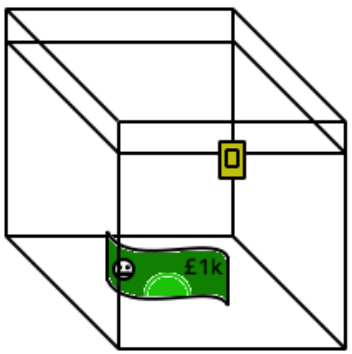
Box B



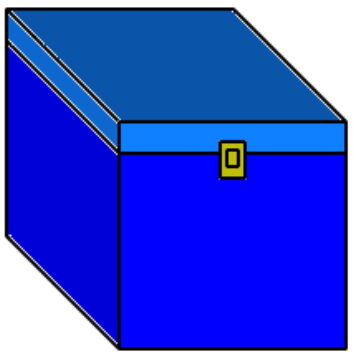
- ...or it contains £1,000,000...

One Box or Two?

Box A



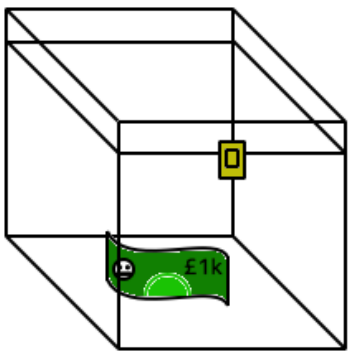
Box B



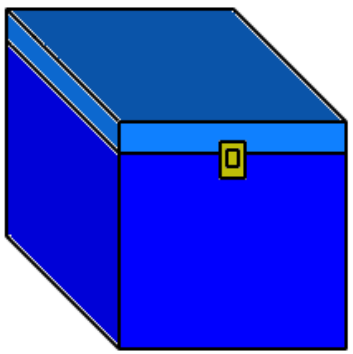
- ...but you do not know which

One Box or Two?

Box A



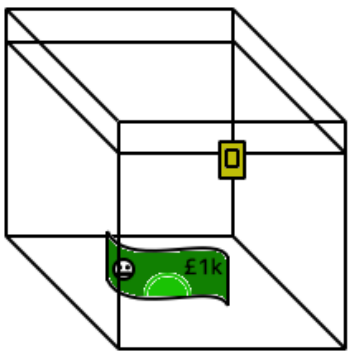
Box B



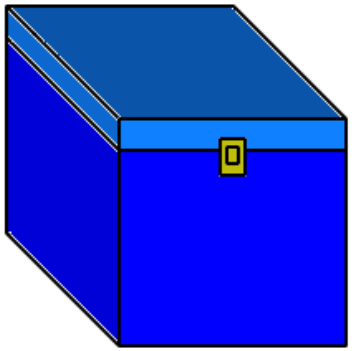
- You are made an offer:

One Box or Two?

Box A



Box B



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

The Predictor

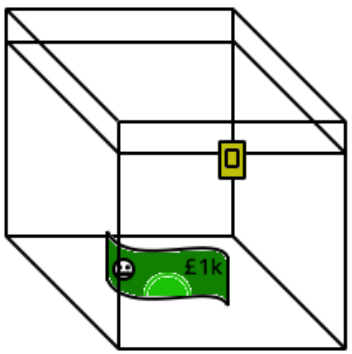
- One week ago, a woman known as the Predictor made a prediction about whether you would take one box or two boxes
- If she predicted that you would only take Box B, she put the £1,000,000 in B
- But if she predicted that you would take both Boxes A and B, she put nothing in B

The Predictor

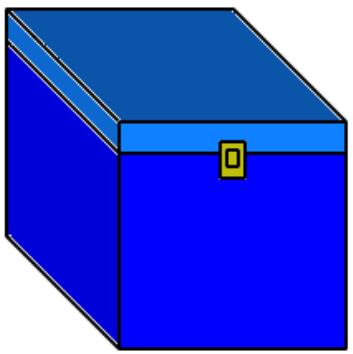
- The Predictor based her prediction on information about you that was available to her last week
- She didn't do it by magically looking into the future or anything like that
- She designed an algorithm which monitors all of your social media activity, and makes a prediction about whether you would take one box or two
- The Predictor's algorithm is **very** reliable
- The Predictor 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 One-Boxing

- If you **two-box** (i.e. take Box A and Box B), then the Predictor will almost certainly have predicted this, and so left Box B empty
- In that case, you'll just get the £1,000 in Box A
- But if you **one-box** (i.e. just take Box B), then the Predictor will almost certainly have predicated this, and so put the money in Box B
- In that case, you'll get £1,000,000
- So you should one-box!!!

An Argument for Two-Boxing

- The Predictor decided whether to put the money in Box B a week ago
- Nothing you do *now* could change what 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
- So you should two-box!!!

Newcomb's Problem

- 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
- This is known as **Newcomb's Problem**
- To solve the problem, we need to figure out which (if either!) argument is sound

Newcomb's Problem

The Argument for Two-Boxing Again

	B is empty	B is not empty
One-box	£0	£1,000,000
Two-box	£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 two-box!

Strict Dominance

- This argument is an application of **Strict Dominance**:
  - $a \succ b$ if: performing a results in a strictly better outcome than performing b in every state
- Two-boxing **strictly dominates** one-boxing

	B is empty	B is not empty
One-box	£0	£1,000,000
Two-box	£1,000	£1,001,000

- In all of the possible states, two-boxing gets you £1,000 more than one-boxing

Problem Solved?

- Strict Dominance seems like a very minimal requirement on rational preference
- All of the decision rules for cases of ignorance that we looked at in Topic 1 imply Strict Dominance
- And Standard Expected Utility validates it too:
  - $EU(a) = [P(s_1) \times U(a \wedge s_1)] + [P(s_2) \times U(a \wedge s_2)]$
  - $EU(b) = [P(s_1) \times U(b \wedge s_1)] + [P(s_2) \times U(b \wedge s_2)]$
  - Therefore, if $U(a \wedge s_1) > U(b \wedge s_1)$ and $U(a \wedge s_2) > U(b \wedge s_2)$, then $EU(a) > EU(b)$
- So problem solved?

Never Study for an Exam!

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

	Pass	Fail
Study	2	0
Not study	3	1

- Not studying dominates studying: whether your pass or fail the exam, you are better off if you did not study
- So by Dominance, 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
- **Dominance totally ignores this fact!**
- So what the "proof" that you should not study for an exam *really* shows is that we have to be very careful about when we apply Dominance

States as Sets of Possible Worlds

- We can think of states as **sets of possible worlds**
  - A possible world is a way that the world could have been
  - There are *lots* of different philosophical accounts of what these worlds really are, but we don't need to get into all of that now

- **EXAMPLES**
  - The state *You pass your exam* is the set of worlds where you pass your exam
  - The state *You fail your exam* is the set of worlds where you fail your exam
  - The state *Box B is empty* is the set of worlds where Box B is empty

Partitioning a State Space

- Suppose we start with a set of possible worlds (a **state space**)
- A **partition** of that state space is just a set of states with the following property:
  - Every world is a member of one of these states, and no world is a member of two of these states
- In general, there are *lots* of ways of partitioning a given state space
  - Partition 1: *You pass your exam*; *You do not pass your exam*
  - Partition 2: *You get a 2i in your exam*; *You do not get a 2i in your exam*

Picking a Partition

- If you want to apply Dominance principles, then you need to choose a partition where all of the states are **independent** of the acts under consideration

	Pass	Fail
Study	2	0
Not study	3	1

- Whether you pass or fail your exam **depends** on whether or not you study

- So we cannot use Dominance on this partition

Depend How?

- If you want to apply Dominance principles, then you need to choose a partition where all of the states are **independent** of the acts under consideration
- What does it mean to say that a state is **independent** of an act?
- There are two different answers to this question, and they lead to two different revisions of Standard Expected Utility Theory
  - **Evidential Decision Theory**
  - **Causal Decision Theory**

Newcomb's Problem

Conditional Probability

- $P(s)$ is the absolute, or **unconditional**, probability of s
- $P(s|a)$ is a **conditional** probability — the probability of s *given* a
- **Informal Gloss:** $P(s|a)$ is the probability you would assign to s if you were working on the assumption of a
- **Formal Definition:** $P(s|a) = P(s \wedge a)/P(a)$

Evidential Decision Theory

- **Standard Expected Utility Theory** uses *unconditional* probabilities:
  - $EU(a) = \sum_{i=1}^{n} P(s_i) \times U(a \wedge s_i)$
- **Evidential Decision Theory** (EDT) uses *conditional* probabilities:
  - $EU_e(a) = \sum_{i=1}^{n} P(s_i|a) \times U(a \wedge s_i)$
- Jeffrey (1965) was the first to present EDT, and Bolker proved a representation theorem for EDT
  - See §3.2 of https://plato.stanford.edu/entries/decision-theory/

Back to Exams

	Pass	Fail
Study	2	0
Not study	3	1

$P(\text{Pass} \mid \text{Study}) = 0.8$; $P(\text{Fail} \mid \text{Study}) = 0.2$

$P(\text{Pass} \mid \neg\text{Study}) = 0.1$; $P(\text{Fail} \mid \neg\text{Study}) = 0.9$

$EU_e(\text{Study}) = [0.8 \times 2] + [0.2 \times 0] = 1.6$ ✓

$EU_e(\neg\text{Study}) = [0.1 \times 3] + [0.9 \times 1] = 1.2$ ×

Back to Dominance

- Dominance applies only when the state space is partitioned into states which are **probabilistically independent** of the acts under consideration

- **Definition:** s is probabilistically independent of a iff: $P(s|a) = P(s)$

Back to Dominance

	Good weather	Bad weather
Fly	2	0
Sail	3	1

$P(G|F) = P(G)$; $P(B|F) = P(B)$
$P(G|S) = P(G)$; $P(B|S) = P(B)$

$EU_e(F) = [P(G) \times 2] + [P(B) \times 0]$ ×
$EU_e(S) = [P(G) \times 3] + [P(B) \times 1]$ ✓

Back to Newcomb

	B is empty	B is not empty
One-box	£0	£1,000,000
Two-box	£1,000	£1,001,000

$P(E|O) = 0.1;\ P(\neg E|O) = 0.9$

$P(E|T) = 0.9;\ P(\neg E|T) = 0.1$

$EU_e(O) = [0.1 \times 0] + [0.9 \times 1,000,000] = 900,000$ ✓

$EU_e(T) = [0.9 \times 1,000] + [0.1 \times 1,001,000] = 101,000$ ×

Problem Solved?

- EDT is a plausible decision theory
  - It allows us to use Dominance reasoning in cases where it seems appropriate...
  - ... and it doesn't force us to use Dominance in cases where it seems inappropriate
- EDT tells us to one-box, so is that the solution to Newcomb's Paradox?
- **No!** There is *another* way of modifying Standard Expected Utility Theory which advocates two-boxing!

Newcomb's Problem

Introducing Causal Dependence

	Pass	Fail
Study	2	0
Not study	3	1

- We cannot apply Dominance here because the states are *dependent* on the acts
- Advocates of EDT say that the kind of dependence which matters is *probabilistic dependence*
- But advocates of **Causal Decision Theory** (CDT) say that it is *causal dependence*

Introducing Causal Dependence

	Pass	Fail
Study	2	0
Not study	3	1

- Your odds of passing or failing your exam are **causally affected** by whether or not you study
- If you study, then that will **cause** your odds of passing to increase
- If you don't study, then that will **cause** your odds of failing to increase

Why Should Causation Matter?

- Decision Theory is meant to tell you how to **act** in various situations
- *Action* is a fundamentally **causal** notion
  - To act in a certain way is, at least in part to cause certain things to happen
- So if we want to figure out whether a given act is rationally preferable, don't we need to focus on its causal consequences?

Causation and Counterfactual Conditionals

- Many philosophers have thought that we could use **counterfactual** (or *subjunctive*) conditionals to analyse causation
  - Pressing the pedal causes the car to accelerate
  - If you were to press the pedal, then the car would accelerate
  - You press the pedal $\Box\!\!\rightarrow$ the car accelerates
- David Lewis developed a detailed theory of counterfactuals, and then developed a detailed analysis of causation in terms of them
  - See volume 2 of his *Collected Papers*
- But for our purposes, we can stick to an intuitive understanding of counterfactuals

Causal Decision Theory

- **Standard Expected Utility Theory** uses *unconditional* probabilities of states:
  - $EU(a) = \sum_{i=1}^{n} P(s_i) \times U(a \wedge s_i)$
- **Evidential Decision Theory** uses *conditional* probabilities of states given acts:
  - $EU_e(a) = \sum_{i=1}^{n} P(s_i|a) \times U(a \wedge s_i)$
- **Causal Decision Theory** uses *unconditional* probabilities of counterfactual conditionals:
  - $EU_c(a) = \sum_{i=1}^{n} P(a \mathrel{\square\!\!\rightarrow} s_i) \times U(a \wedge s_i)$

    (This is Gibbard and Harper's version of CDT, but there are lots of others. Lewis develops his own in his (1981), and compares it to other versions in §§6–9)

Back to Exams (*again*)

	Pass	Fail
Study	2	0
Not study	3	1

$P(\text{Study} \boxdot\!\!\rightarrow \text{Pass}) = 0.75$

$P(\text{Study} \boxdot\!\!\rightarrow \text{Fail}) = 0.25$

$P(\neg\text{Study} \boxdot\!\!\rightarrow \text{Pass}) = 0.1$

$P(\neg\text{Study} \boxdot\!\!\rightarrow \text{Fail}) = 0.9$

$EU_c(\text{Study}) = [0.75 \times 2] + [0.25 \times 0] = 1.5$ ✓

$EU_c(\neg\text{Study}) = [0.1 \times 3] + [0.9 \times 1] = 1.2$ ×

Back to Dominance (*again*)

- Dominance applies only when the state space is partitioned into states which are **causally independent** of the acts under consideration

- **Definition:** The states of nature in a given partition are causally independent of the acts in a given alternative set iff every state of nature in that partition, s, and every pair of acts in that alternative set, a and b, meet this condition:
  - $P(a \mathrel{\square\!\!\rightarrow} s) = P(b \mathrel{\square\!\!\rightarrow} s)$

Back to Dominance (*again*)

	Good weather	Bad weather
Fly	2	0
Sail	3	1

$$P(F \boxrightarrow G) = P(S \boxrightarrow G)$$
$$P(F \boxrightarrow B) = P(S \boxrightarrow B)$$

$$EU_c(F) = [P(F \boxrightarrow G) \times 2] + [P(F \boxrightarrow B) \times 0] \quad \times$$
$$EU_c(S) = [P(S \boxrightarrow G) \times 3] + [P(S \boxrightarrow B) \times 1] \quad \checkmark$$

Back to Newcomb (*again*)

	B is empty	B is not empty
One-box	£0	£1,000,000
Two-box	£1,000	£1,001,000

$$P(O \boxrightarrow E) = P(T \boxrightarrow E)$$

$$P(O \boxrightarrow \neg E) = P(T \boxrightarrow \neg E)$$

Whatever you do now, whether B is full or empty is already fixed and settled!

$$EU_c(O) = [P(O \boxrightarrow E) \times 0] + [P(O \boxrightarrow \neg E) \times 1,000,000] \quad \times$$

$$EU_c(T) = [P(T \boxrightarrow E) \times 1,000] + [P(T \boxrightarrow \neg E) \times 1,001,000] \quad \checkmark$$

EDT versus CDT

- EDT and CDT are both improvements on Standard Expected Utility Theory
- In most everyday circumstances, they even agree on their recommendations!
  - In most normal circumstances, if $P(s|a) > P(s)$, then that is only because $P(a \mathrel{\square\!\!\rightarrow} s)$ is high
- But Newcomb's Problem shows that they do not *always* agree
  - EDT recommends one-boxing, but CDT recommends two-boxing
- In the next lecture, we will look at the reasons for preferring one of these decision theories over the other

References

- Gibbard, Allan and Harper, William (1978) 'Counterfactuals and Two Kinds of Expected Utility', in Michael J. Hooker et al eds, *Foundations and Applications of Decision Theory*, pp. 125–162, Dordrecht: D. Reidel Publishing Co.
- Jeffrey, Richard (1965) *The Logic of Decision*, New York, NY: McGraw-Hill Book Co.
- Lewis, David (1981) 'Causal Decision Theory', *Australasian Journal of Philosophy* 59: 5–30
- —— (1987) *Collected Papers* vol. 2, Oxford: Oxford University Press