Rationality, Morality and Economics

Topic 3, Lecture 2

Newcomb's Problem

Rob Trueman
rob.trueman@york.ac.uk

University of York

Newcomb's Problem

Three Decision Theories

- **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** (EDT) 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** (CDT) uses *unconditional* probabilities of counterfactual conditionals:
  - $EU_c(a) = \sum_{i=1}^{n} P(a \mathrel{\Box\!\!\rightarrow} s_i) \times U(a \wedge s_i)$

EDT or CDT?

- EDT and CDT make the same recommendations in most "real life" decision problems
- In most normal circumstances, if $P(s|a) > P(s)$, then that is only because $P(a \mathrel{\Box\!\!\rightarrow} s)$ is high
- But they make different recommendations in the **Newcomb Problem**

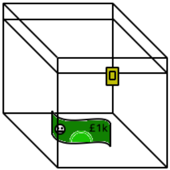
One Box or Two?

Box A Box B

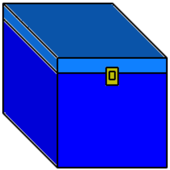
- You are presented with two boxes

One Box or Two?

Box A



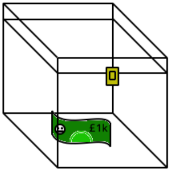
Box B



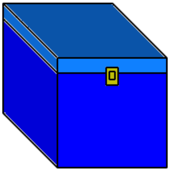
- Box A is transparent, and you can see that it contains £1,000

One Box or Two?

Box A



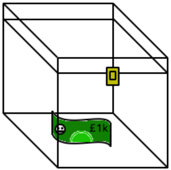
Box B



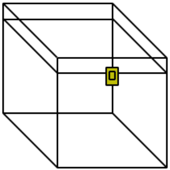
- Box B is opaque, and you cannot see what is in it

One Box or Two?

Box A



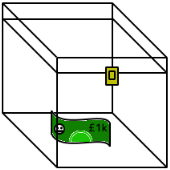
Box B



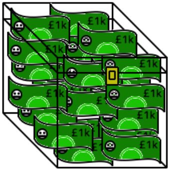
- You know that Box B is either empty...

One Box or Two?

Box A



Box B



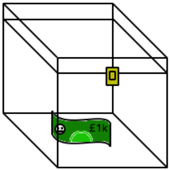
- ...or it contains £1,000,000...

One Box or Two?

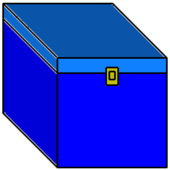
- ...but you do not know which

One Box or Two?

Box A



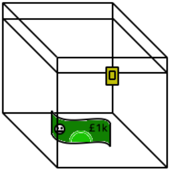
Box B



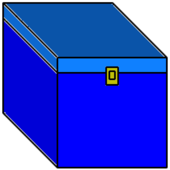
- You are made an offer:

One Box or Two?

Box A



Box B



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

The Predictor

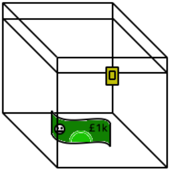
- One week ago, a woman known as the Predictor made a prediction about whether you would take one box or two boxes
- If she predicted that you would only take Box B, she put the £1,000,000 in B
- But if she predicted that you would take both Boxes A and B, she put nothing in B

The Predictor

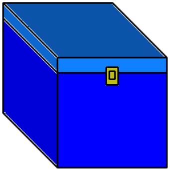
- The Predictor based her prediction on information about you that was available to her last week
- The Predictor's is **very** reliable
- The Predictor has played this game with lots and lots of people, and her predictions have always been right

One Box or Two?

Box A



Box B



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

One-Boxing is E-Rational

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

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

$EU_e(O) = [0.1 \times 0] + [0.9 \times 1,000,000] = 900,000$ ✓
$EU_e(T) = [0.9 \times 1,000] + [0.1 \times 1,001,000] = 101,000$ ×

Two-Boxing is C-Rational

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

$$P(O \Box\!\!\rightarrow E) = P(T \Box\!\!\rightarrow E)$$

$$P(O \Box\!\!\rightarrow \neg E) = P(T \Box\!\!\rightarrow \neg E)$$

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

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

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

Newcomb's Problem

An Argument for EDT

- Everyone agrees that if you take EDT's advice and one-box, then you will probably get £$1,000,000$
- Everyone agrees that if you take CDT's advice and two-box, then you will probably get only £$1,000$
- So isn't it obvious that EDT is the right view of rationality?
  - Following EDT predictably gets you more of what you value
- **The Challenge to CDT:** *Why ain'cha rich?*

Lewis's Defence of CDT

- **Lewis's (1981b) Answer:** It was never an option for me to get rich!
- A two-boxer takes all the money that's available to them in the Newcomb problem, it's just that there's only £1,000 in the two boxes
- The two-boxer's choice to take both boxes didn't deprive them of any money
- It **maximised** the amount of money that they could get out of the situation they were confronted with!

A Point in EDT's Favour

- According to Lewis, Newcomb's Problem is generated by the Predictor's decision to reward people who will *irrationally* one-box
- However, Lewis also recognises that there is an important asymmetry between CDT and EDT
- The standard Newcomb problem where people are rewarded for being C-irrational is **logically coherent**
- A Newcomb-style problem where people are rewarded for being E-irrational would be **logically incoherent**

A Point in EDT's Favour

- Imagine I told you that the Predictor would put £$1,000,000$ into Box B iff you two-box
- In that case, two-boxing is E-rational!
- More generally, if we try to set-up a Newcomb-style problem where the Predictor rewards a certain choice, she automatically makes it the **E-rational** choice
- So it is impossible for her to reward **E-irrational** choices

Newcomb's Problem

A Bit Too Sci-Fi?

- Newcomb's Problem is so unrealistic that you might not think it can tell us anything very interesting about rationality
  - If a decision theory gets it wrong in Newcomb's Problem, who cares?
  - Can we even rely on our intuitions to tell us what is the right decision in Newcomb's Problem?
- Philosophers hae tried to deal with this problem by finding more realistic versions of Newcomb's Problem
- And notably, the equivalent of two-boxing is generally agreed to be the rational course of action in these realistic Newcomb Problems

A Medical Problem

- 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 Smoke?

- This problem is structurally identical to the Newcomb Problem
  - If you find yourself smoking, then that should increase your credence that you have cancer
  - But smoking doesn't **cause** cancer the gene does, and even if you force yourself not to smoke, you will still have the gene

- But this problem is a lot more realistic than the traditional Newcomb Problem
  - It doesn't really have to be **true** that the gene causes smoking and cancer
  - Since EDT and CDT *both* use subjective credences to calculate expected utility, all that matters is that the agent in the problem **believes** that it is true

What EDT Says

	Cancer	No Cancer
Smoke	-100	50
Don't Smoke	-150	0

$P(C|S) = 0.9; P(\neg C|S) = 0.1$
$P(C|\neg S) = 0.2; P(\neg C|\neg S) = 0.8$

$EU_e(S) = [0.9 \times -100] + [0.1 \times 50] = -85$ ×
$EU_e(\neg S) = [0.2 \times -150] + [0.8 \times 0] = -30$ ✓

What CDT Says

	Cancer	No Cancer
Smoke	-100	50
Don't Smoke	-150	0

$$P(S \mathrel{\Box\!\!\rightarrow} C) = P(\neg S \mathrel{\Box\!\!\rightarrow} C)$$
$$P(S \mathrel{\Box\!\!\rightarrow} \neg C) = P(\neg S \mathrel{\Box\!\!\rightarrow} \neg C)$$

$$EU_c(S) = [P(S \mathrel{\Box\!\!\rightarrow} C) \times -100] + [P(S \mathrel{\Box\!\!\rightarrow} \neg C) \times 50] \quad \checkmark$$
$$EU_c(\neg S) = [P(\neg S \mathrel{\Box\!\!\rightarrow} C) \times -150] + [P(\neg S \mathrel{\Box\!\!\rightarrow} \neg C) \times 0] \quad \times$$

A Victory for CDT?

- Most people think that it is obviously irrational to quit smoking in the Medical Newcomb Problem
  - Not smoking now cannot change whether you have the gene which causes cancer
  - But not smoking now will rob you of the pleasure of smoking
- So does that show CDT is right and EDT is wrong?
- *Not yet!* A number of EDTers have argued that EDT actually recommends that you smoke in the Medical Newcomb Problem

Newcomb's Problem

How does the Gene Cause Smoking?

(1) **The gene makes you want to smoke**

- This is the only plausible and relevant explanation

(2) **The gene makes you compulsively smoke against your will**

- If you aren't really choosing whether to smoke or not, then we can't really dicuss whether your choices are rational

(3) **Magic**

- The Medical Newcomb Problem is meant to be *more* realistic than the traditional Newcomb Problem!

Knowing Your Own Mind

- A fully rational agent should be aware of their own beliefs and desires
- So a fully rational agent should be aware if they want to smoke
- If they do notice that they want to smoke, then that should increase their credence that they have the smoking gene, and so increase their credence that they will get cancer
- But once that has happened, their credence shouldn't be *further* affected by whether or not they actually go on to smoke

A Comparisson

- Imagine a car drives past, and that you have never seen that car before
- Seeing this should obviously increase your credence that someone turned the ignition key in that car
- But it shouldn't affect your credence if you could already hear the car's engine running
  - Hearing the engine running and seeing it drive past are both evidence that the ignition key was turned
  - But once you have one of these pieces of evidence, getting the other shouldn't boost your credence that the key was turned

A Comparisson

- In the Medical Newcomb Problem, feeling a desire to smoke and actually smoking are both evidence that you have the smoking gene
- But once you have one of these pieces of evidence, getting the other shouldn't boost your credence that you have the gene
- If you are maximally rational, you already know your desires, and so already know if you have the desire to smoke
- So whether you actually smoke shouldn't affect your credences

Updating Your Credences

- In the Medical Newcomb Problem, you should first check whether you want to smoke, and update your credence that you will get cancer
  - If you do want to smoke (W), then set your credence as: $P(C) = P(C|W)$
  - If you don't want to smoke ($\neg W$), then set your credence as: $P(C) = P(C|\neg W)$

- Once you know whether or not you want to smoke, actually smoking doesn't change your credences at all
  - $P(C|W \wedge S) = P(C|W \wedge \neg S) = P(C|W)$
  - $P(C|\neg W \wedge S) = P(C|\neg W \wedge \neg S) = P(C|\neg W)$
  - Therefore, $P(C|S) = P(C|\neg S) = P(C)$

Updating Your Credences

- In the Medical Newcomb Problem, you should first check whether you want to smoke, and update your credence that you will get cancer
  - If you do want to smoke (W), then set your credence as: $P(C) = P(C|W)$
  - ~~If you don't want to smoke ($\neg W$), then set your credence as: $P(C) = P(C|\neg W)$~~

- Once you know whether or not you want to smoke, actually smoking doesn't change your credences at all
  - $P(C|W \wedge S) = P(C|W \wedge \neg S) = P(C|W)$
  - $\cancel{P(C|\neg W \wedge S) = P(C|\neg W \wedge \neg S) = P(C|\neg W)}$
  - Therefore, $P(C|S) = P(C|\neg S) = P(C)$

Updating Your Credences

- In the Medical Newcomb Problem, you should first check whether you want to smoke, and update your credence that you will get cancer
  - If you do want to smoke (W), then set your credence as: $P(C) = P(C|W)$
  - ~~If you don't want to smoke ($\neg W$), then set your credence as:~~ $\cancel{P(C) = P(C|\neg W)}$

- Once you know whether or not you want to smoke, actually smoking doesn't change your credences at all
  - $P(C|S) = P(C|\neg S) = P(C)$
  - $\cancel{P(C|\neg W \wedge S) = P(C|\neg W \wedge \neg S) = P(C|\neg W)}$
  - Therefore, $P(C|S) = P(C|\neg S) = P(C)$

Updating Your Credences

- In the Medical Newcomb Problem, you should first check whether you want to smoke, and update your credence that you will get cancer
  - ~~If you do want to smoke (W), then set your credence as: $P(C) = P(C|W)$~~
  - If you don't want to smoke ($\neg W$), then set your credence as: $P(C) = P(C|\neg W)$

- Once you know whether or not you want to smoke, actually smoking doesn't change your credences at all
  - $\cancel{P(C|W \wedge S) = P(C|W \wedge \neg S) = P(C|W)}$
  - $P(C|\neg W \wedge S) = P(C|\neg W \wedge \neg S) = P(C|\neg W)$
  - Therefore, $P(C|S) = P(C|\neg S) = P(C)$

Updating Your Credences

- In the Medical Newcomb Problem, you should first check whether you want to smoke, and update your credence that you will get cancer
  - ~~If you do want to smoke (W), then set your credence as:~~ $\cancel{P(C) = P(C|W)}$
  - If you don't want to smoke ($\neg W$), then set your credence as: $P(C) = P(C|\neg W)$

- Once you know whether or not you want to smoke, actually smoking doesn't change your credences at all
  - $\cancel{P(C|W \wedge S) = P(C|W \wedge \neg S) = P(C|W)}$
  - $P(C|S) = P(C|\neg S) = P(C)$
  - Therefore, $P(C|S) = P(C|\neg S) = P(C)$

Updating Your Credences

- In the Medical Newcomb Problem, you should first check whether you want to smoke, and update your credence that you will get cancer
  - If you do want to smoke (W), then set your credence as: $P(C) = P(C|W)$
  - If you don't want to smoke ($\neg W$), then set your credence as: $P(C) = P(C|\neg W)$
- Once you know whether or not you want to smoke, actually smoking doesn't change your credences at all
  - $P(C|W \wedge S) = P(C|W \wedge \neg S) = P(C|W)$
  - $P(C|\neg W \wedge S) = P(C|\neg W \wedge \neg S) = P(C|\neg W)$
  - Therefore, $P(C|S) = P(C|\neg S) = P(C)$

Updating Your Credences

- In the Medical Newcomb Problem, you should first check whether you want to smoke, and update your credence that you will get cancer
  - If you do want to smoke (W), then set your credence as: $P(C) = P(C|W)$
  - If you don't want to smoke ($\neg W$), then set your credence as: $P(C) = P(C|\neg W)$

- Once you know whether or not you want to smoke, actually smoking doesn't change your credences at all
  - $P(C|W \wedge S) = P(C|W \wedge \neg S) = P(C|W)$
  - $P(C|\neg W \wedge S) = P(C|\neg W \wedge \neg S) = P(C|\neg W)$
  - Therefore, $P(C|S) = P(C|\neg S) = P(C)$
  - It also follows that $P(\neg C|S) = P(\neg C|\neg S) = P(\neg C)$

What EDT Says Now

	Cancer	No Cancer
Smoke	-100	50
Don't Smoke	-150	0

$$P(C|S) = P(C|\neg S) = P(C)$$
$$P(\neg C|S) = P(\neg C|\neg S) = P(C)$$

$$EU_e(S) = [P(C) \times -100] + [P(\neg C) \times 50] \quad \checkmark$$
$$EU_e(\neg S) = [P(C) \times -150] + [P(\neg C) \times 0] \quad \times$$

Lewis's Objection

- The Tickle Defence relies on the assumption that a fully rational agent should know all of their beliefs and desires
- Lewis (1981a: 10–11) objected that while this might be fine for **fully rational** agents, real agents are not like that
- So the Tickle Defence is useless for merely **partly rational** agents like us
- Agents like *us* should use CDT, not EDT

Responding to Lewis

> *One should not object here that a person's desires may not always be accessible to introspection. This is true but irrelevant. Our [Tickle Defence] needs to be employed only for situations that provide alleged counterexamples to [EDT]. And there can be a counterexample to [EDT] only if the [theory] is applied, and therefore only if the beliefs and desires of the agent are known by him at the time of deliberation.*
>
> *(Horwich 1987: 183)*

Newcomb's Problem

Other Realistic Problems

- A number of "real life" Newcomb Problems have been discussed
- Most interestingly, Lewis (1979) argued that the classic Prisonner's Dilemma is a version of the Newcomb Problem
- It is an open question whether the Tickle Defence can be used to undermine all of these other Newcomb-style Problems
  - For attempts to use the Tickle Defence in a range of cases, see: Horwich 1981: ch. 11; Ahmed 2014: ch. 4

Ratifiability?

- Some EDTers have tried to find different ways of defending their theory
- Jeffrey (1981) suggested tweaking EDT by insisting that a rational decision must be **ratifiable**
- According to this idea, act A is rational only if there is no act B such that the value of B exceeds the value of A on the supposition that A is the act decided upon
  - Not smoking in the Medical Newcomb Problem is unratifiable
  - Once you choose not to smoke, whether you actually smoke ceases to serve as evidence that you have the bad gene
  - At that point, smoking becomes E-rational!

Ratifiability?

- Egan (2007: 107–13) argues that insisting that rational decisions must be ratifiable cannot save CDT or EDT
- In fact, he goes even further: he thinks that *nothing* can save CDT or EDT
- We will discuss Egan's paper in the seminar

References

- Ahmed, Arif (2014) *Evidence, Decision and Causality*, Cambridge: CUP
- Egan, Andy (2007) 'Some Counterexamples to Causal Decision Theory', *The Philosophical Review* 116: 93–114
- Horwich, Paul (1987) *Asymmetries in Time*, Cambridge, MA: MIT Press
- Jeffrey, Richard (1981) '*The Logic of Decision* Defended', *Synthese* 48: 473–492
- Lewis, David (1979) 'Prisoners' Dilemma is a Newcomb Problem', *Philosophy & Public Affairs* 8: 235–240
- —— (1981a) 'Causal Decision Theory', *Australasian Journal of Philosophy* 59: 5–30
- —— (1981b) 'Why Ain'cha Rich?', *Noûs* 15: 377–380