Decision and Social Choice

PE3101P — Lecture 3: Newcomb's problem

Tomi Francis

National University of Singapore

27th August, 2026

Today

1. Recap
2. More Probability
3. Newcomb's problem
4. EDT and CDT
5. EDT vs CDT

First For-Credit Assignment

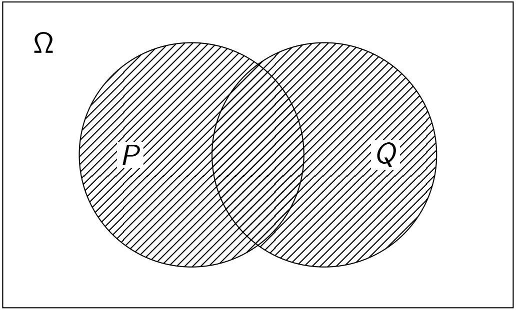
- Goes up on my website soon.
- Due **Tuesday of Week 5 — 8$^{\text{th}}$ September**, by 23:59.
- Submission still TBD but most likely on Canvas.
- Lecture slides up after the lecture. (Proper lecture notes hopefully to follow soon.)
- Subject to change until Saturday.

Sample Spaces and Events

- A **sample space** Ω: all the ways things might turn out.
- An **event**: a subset of Ω — the outcomes at which it happens.
- A **credence function** Cr: a function from events to real numbers. If it's a nice credence function, it satisfied the probability axioms.

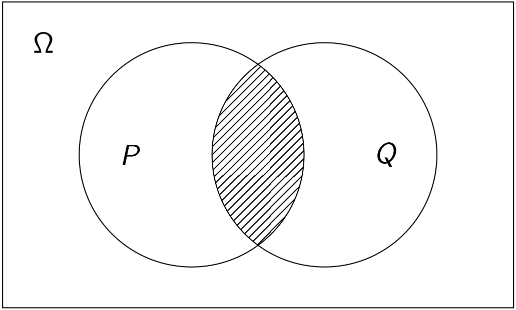
Combining Events

$P \cup Q$ — **either** happens



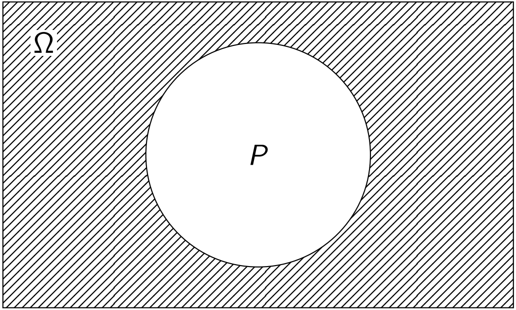
Combining Events

$P \cap Q$ — **both** happen



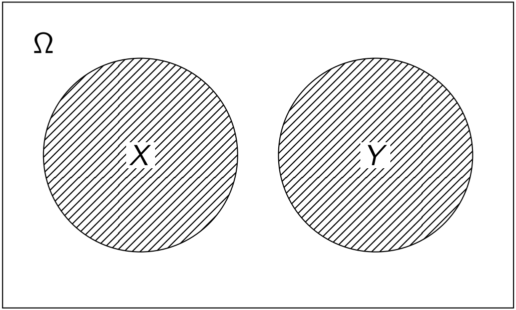
Combining Events

P^c — everything in Ω **outside** P



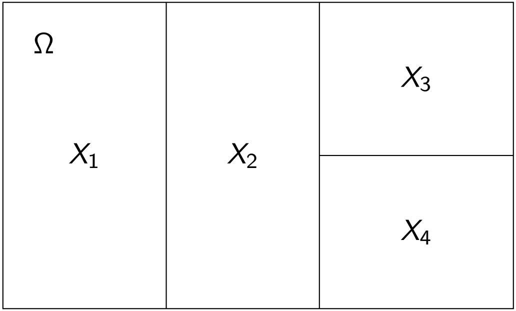
Mutual Exclusivity and Joint Exhaustiveness

Mutually exclusive: $X \cap Y = \varnothing$. At most one happens.



Mutual Exclusivity and Joint Exhaustiveness

Jointly exhaustive as well: they **partition** Ω. Exactly one happens.



Which is what we want from a set of states of nature.

The Probability Axioms

(i) **Non-negativity**: $\mathrm{Cr}(\sigma) \geq 0$.

(ii) **Normalisation**: $\mathrm{Cr}(\Omega) = 1$.

(iii) **Finite additivity**: if $X \cap Y = \varnothing$, then $\mathrm{Cr}(X \cup Y) = \mathrm{Cr}(X) + \mathrm{Cr}(Y)$.

Conditional Probabilities and Conditionalisation

Definition

$$\mathrm{Cr}(A \mid B) = \frac{\mathrm{Cr}(A \cap B)}{\mathrm{Cr}(B)} \qquad \text{provided } \mathrm{Cr}(B) > 0.$$

- Areas are probabilities.
- Conditioning on B **throws away everything outside** B, and rescales what's left.

Conditional Probabilities and Conditionalisation

Definition

$$\mathrm{Cr}(A \mid B) = \frac{\mathrm{Cr}(A \cap B)}{\mathrm{Cr}(B)} \qquad \text{provided } \mathrm{Cr}(B) > 0.$$

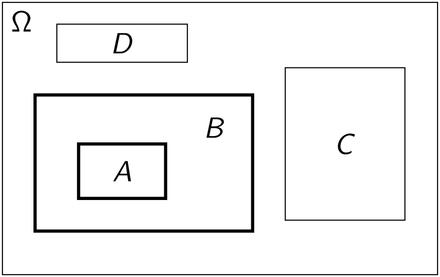
- Areas are probabilities.
- Conditioning on B **throws away everything outside** B, and rescales what's left.

The principle of conditionalisation

Learn E and nothing more, and your new credence in A should be

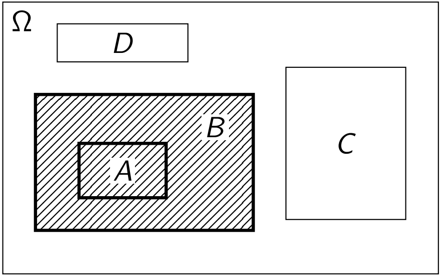
$$\mathrm{Cr}_{\text{new}}(A) = \mathrm{Cr}(A \mid E).$$

Conditionalisation, in Pictures



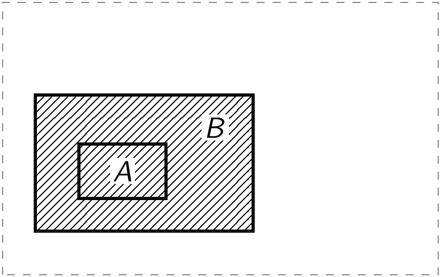
Areas are probabilities. B is what we'll condition on, and A lies wholly inside it.

Conditionalisation, in Pictures



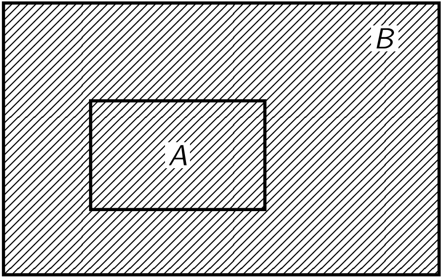
Shade the conditioning event B.

Conditionalisation, in Pictures



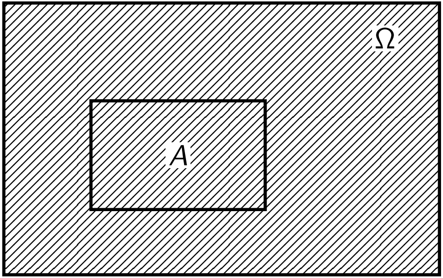
Discard everything outside B. **Nothing has moved or changed size yet.**

Conditionalisation, in Pictures



Blow B up until it fills the space Ω used to. The scaling is uniform, so every ratio of areas inside B is unchanged.

Conditionalisation, in Pictures



Relabel: B **is the new sample space**, and $\mathrm{Cr}(A \mid B)$ is the fraction of it that A occupies.

Probabilistic Independence

Definition

$$\mathrm{Cr}(A \cap B) = \mathrm{Cr}(A) \cdot \mathrm{Cr}(B)$$

or equivalently, where $\mathrm{Cr}(B) > 0$,

$$\mathrm{Cr}(A \mid B) = \mathrm{Cr}(A).$$

- Learning whether B happened makes no difference at all to how likely A is.

The Law of Total Probability

The Law of Total Probability

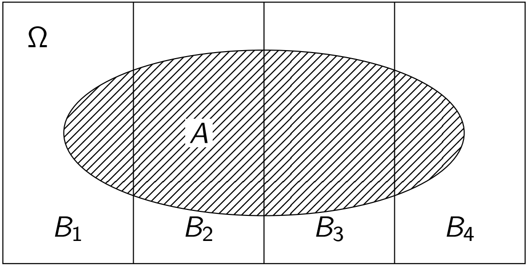
Where $B_1, \ldots, B_n$ partition Ω:

$$\mathrm{Cr}(A) = \mathrm{Cr}(B_1)\,\mathrm{Cr}(A \mid B_1) + \cdots + \mathrm{Cr}(B_n)\,\mathrm{Cr}(A \mid B_n).$$

The Law of Total Probability

Where $B_1, \ldots, B_n$ partition Ω:

$$Cr(A) = Cr(B_1)\,Cr(A \mid B_1) + \cdots + Cr(B_n)\,Cr(A \mid B_n).$$

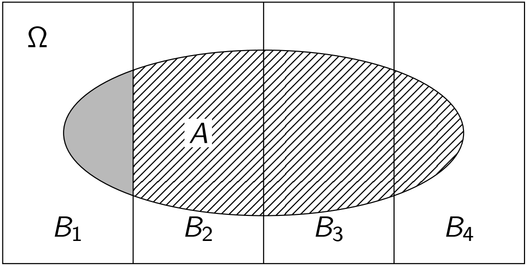


- The B_i cut A into the disjoint pieces $A \cap B_i$.

The Law of Total Probability

Where $B_1, \ldots, B_n$ partition Ω:

$$\mathrm{Cr}(A) = \mathrm{Cr}(B_1)\,\mathrm{Cr}(A \mid B_1) + \cdots + \mathrm{Cr}(B_n)\,\mathrm{Cr}(A \mid B_n).$$

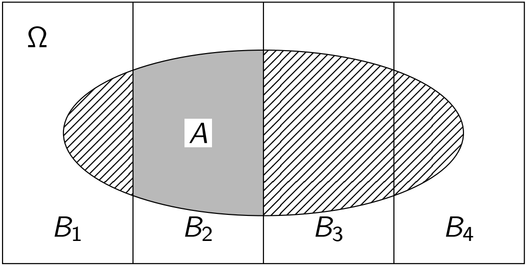


- The B_i cut A into the disjoint pieces $A \cap B_i$.

The Law of Total Probability

Where $B_1, \ldots, B_n$ partition Ω:

$$\mathrm{Cr}(A) = \mathrm{Cr}(B_1)\,\mathrm{Cr}(A \mid B_1) + \cdots + \mathrm{Cr}(B_n)\,\mathrm{Cr}(A \mid B_n).$$

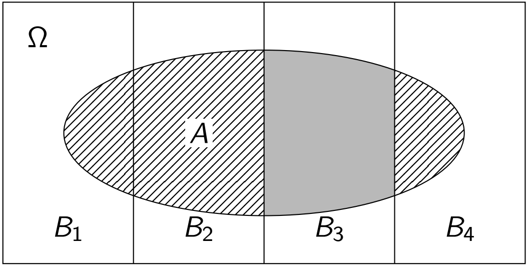


- The B_i cut A into the disjoint pieces $A \cap B_i$.

The Law of Total Probability

Where $B_1, \ldots, B_n$ partition Ω:

$$\mathrm{Cr}(A) = \mathrm{Cr}(B_1)\,\mathrm{Cr}(A \mid B_1) + \cdots + \mathrm{Cr}(B_n)\,\mathrm{Cr}(A \mid B_n).$$

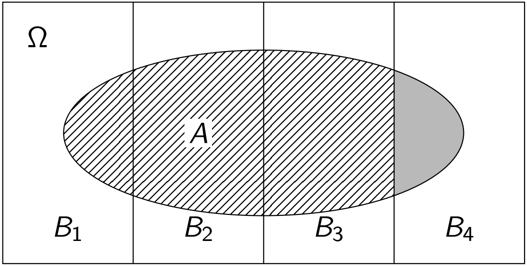


- The B_i cut A into the disjoint pieces $A \cap B_i$.

The Law of Total Probability

Where $B_1, \ldots, B_n$ partition Ω:

$$\mathrm{Cr}(A) = \mathrm{Cr}(B_1)\,\mathrm{Cr}(A \mid B_1) + \cdots + \mathrm{Cr}(B_n)\,\mathrm{Cr}(A \mid B_n).$$

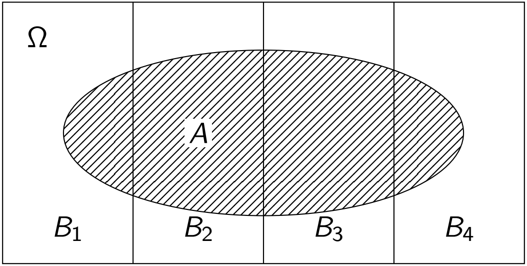


- The B_i cut A into the disjoint pieces $A \cap B_i$.

The Law of Total Probability

Where $B_1, \ldots, B_n$ partition Ω:

$$\mathrm{Cr}(A) = \mathrm{Cr}(B_1)\,\mathrm{Cr}(A \mid B_1) + \cdots + \mathrm{Cr}(B_n)\,\mathrm{Cr}(A \mid B_n).$$

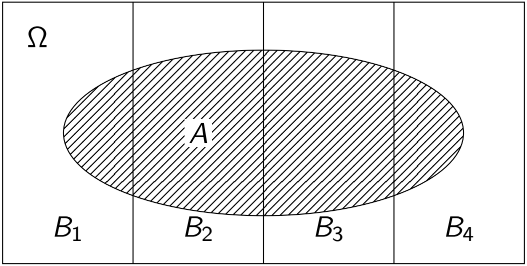


- So by finite additivity, $\mathrm{Cr}(A) = \mathrm{Cr}(A \cap B_1) + \cdots + \mathrm{Cr}(A \cap B_n)$.

The Law of Total Probability

Where $B_1, \ldots, B_n$ partition Ω:

$$\mathrm{Cr}(A) = \mathrm{Cr}(B_1)\,\mathrm{Cr}(A \mid B_1) + \cdots + \mathrm{Cr}(B_n)\,\mathrm{Cr}(A \mid B_n).$$



- And $\mathrm{Cr}(A \cap B_i) = \mathrm{Cr}(B_i)\,\mathrm{Cr}(A \mid B_i)$, by the definition. Substitute.

Bayes' Theorem

Bayes' Theorem

Bayes' theorem

$$\mathrm{Cr}(A \mid B) = \frac{\mathrm{Cr}(B \mid A)\ \mathrm{Cr}(A)}{\mathrm{Cr}(B)} \qquad \text{provided } \ \mathrm{Cr}(A), \mathrm{Cr}(B) > 0.$$

Proving Bayes' Theorem

$$\text{Cr}(A \mid B) = \frac{\text{Cr}(B \mid A)\,\text{Cr}(A)}{\text{Cr}(B)}$$

What we want to show.

Proving Bayes' Theorem

$$\frac{\mathbf{Cr}(\boldsymbol{A \cap B})}{\mathbf{Cr}(\boldsymbol{B})} = \frac{\mathrm{Cr}(B \mid A)\,\mathrm{Cr}(A)}{\mathrm{Cr}(B)}$$

Substitute the definition of $\mathrm{Cr}(A \mid B)$:

$$\mathrm{Cr}(A \mid B) = \mathrm{Cr}(A \cap B) / \mathrm{Cr}(B)$$

Proving Bayes' Theorem

$$\frac{\mathrm{Cr}(A \cap B)}{\mathrm{Cr}(B)} = \frac{\left(\mathbf{Cr}(\boldsymbol{A} \cap \boldsymbol{B}) / \mathbf{Cr}(\boldsymbol{A})\right) \mathrm{Cr}(A)}{\mathrm{Cr}(B)}$$

Substitute the definition of $\mathrm{Cr}(B \mid A)$:

$$\mathrm{Cr}(B \mid A) = \mathrm{Cr}(A \cap B) / \mathrm{Cr}(A)$$

Proving Bayes' Theorem

$$\frac{\mathrm{Cr}(A \cap B)}{\mathrm{Cr}(B)} = \frac{\mathrm{Cr}(A \cap B)\ \mathrm{Cr}(A)}{\mathbf{Cr(\boldsymbol{A})}\ \mathrm{Cr}(B)}$$

$\mathrm{Cr}(A)$ was dividing the numerator, so it belongs in the denominator.

Proving Bayes' Theorem

$$\frac{Cr(A \cap B)}{Cr(B)} = \frac{Cr(A \cap B)}{Cr(B)}$$

$Cr(A)$ appears top and bottom, so it cancels.

Proving Bayes' Theorem

$$\frac{\mathrm{Cr}(A \cap B)}{\mathrm{Cr}(B)} = \frac{\mathrm{Cr}(A \cap B)}{\mathrm{Cr}(B)}$$

The two sides are now identical.

Bayes Factors

- Two hypotheses H_1, H_2, and evidence E arrives.

Bayes Factors

- Two hypotheses H_1, H_2, and evidence E arrives.
- The **Bayes factor** is $\mathrm{Cr}(E \mid H_1) / \mathrm{Cr}(E \mid H_2)$.

Bayes Factors

- Two hypotheses H_1, H_2, and evidence E arrives.
- The **Bayes factor** is $\mathrm{Cr}(E \mid H_1)/\,\mathrm{Cr}(E \mid H_2)$.
- How much better E is explained by H_1 than by H_2.

In Odds Form

$$\underbrace{\frac{\mathrm{Cr}(H_1 \mid E)}{\mathrm{Cr}(H_2 \mid E)}}_{\text{posterior odds}} = \underbrace{\frac{\mathrm{Cr}(E \mid H_1)}{\mathrm{Cr}(E \mid H_2)}}_{\text{Bayes factor}} \times \underbrace{\frac{\mathrm{Cr}(H_1)}{\mathrm{Cr}(H_2)}}_{\text{prior odds}}$$

Proving the Odds Form

$$\frac{\mathrm{Cr}(H_1 \mid E)}{\mathrm{Cr}(H_2 \mid E)}$$

The posterior odds.

Proving the Odds Form

$$\frac{\left(\mathbf{Cr}(\boldsymbol{E} \mid \boldsymbol{H_1})\ \mathbf{Cr}(\boldsymbol{H_1}) / \mathbf{Cr}(\boldsymbol{E})\right)}{\mathrm{Cr}(H_2 \mid E)}$$

Substitute Bayes' theorem for $\mathrm{Cr}(H_1 \mid E)$:

$$\mathrm{Cr}(H_1 \mid E) = \mathrm{Cr}(E \mid H_1)\ \mathrm{Cr}(H_1) / \mathrm{Cr}(E)$$

Proving the Odds Form

$$\frac{\mathrm{Cr}(E \mid H_1)\ \mathrm{Cr}(H_1)}{\boldsymbol{\mathrm{Cr}(E)}\ \mathrm{Cr}(H_2 \mid E)}$$

$\mathrm{Cr}(E)$ was dividing the numerator, so it belongs in the denominator.

Proving the Odds Form

$$\frac{\text{Cr}(E \mid H_1)\ \text{Cr}(H_1)}{\text{Cr}(E)\left(\boldsymbol{\text{Cr}(E \mid H_2)\ \text{Cr}(H_2)/\ \text{Cr}(E)}\right)}$$

Substitute Bayes' theorem for $\text{Cr}(H_2 \mid E)$:

$$\text{Cr}(H_2 \mid E) = \text{Cr}(E \mid H_2)\ \text{Cr}(H_2)/\ \text{Cr}(E)$$

Proving the Odds Form

$$\frac{\mathrm{Cr}(E \mid H_1)\ \mathrm{Cr}(H_1)\ \boldsymbol{\mathrm{Cr}(E)}}{\mathrm{Cr}(E)\ \mathrm{Cr}(E \mid H_2)\ \mathrm{Cr}(H_2)}$$

This $\mathrm{Cr}(E)$ was dividing the denominator, so it belongs in the numerator.

Proving the Odds Form

$$\frac{\mathrm{Cr}(E \mid H_1)\ \mathrm{Cr}(H_1)}{\mathrm{Cr}(E \mid H_2)\ \mathrm{Cr}(H_2)}$$

$\mathrm{Cr}(E)$ now appears top and bottom, so it cancels.

Proving the Odds Form

$$\frac{Cr(E \mid H_1)}{Cr(E \mid H_2)} \times \frac{Cr(H_1)}{Cr(H_2)}$$

Separate the two conditional credences from the two priors.

Proving the Odds Form

$$\frac{\mathrm{Cr}(E \mid H_1)}{\mathrm{Cr}(E \mid H_2)} \times \frac{\mathrm{Cr}(H_1)}{\mathrm{Cr}(H_2)}$$

Bayes factor $\times$ prior odds.

Dutch Books

The betting interpretation

An agent whose credence in A is p will buy a $\$S$ bet on A for any price below $\$pS$, and sell one for any price above $\$pS$.

Dutch Books

The betting interpretation

An agent whose credence in A is p will buy a \$$S$ bet on A for any price below \$$pS$, and sell one for any price above \$$pS$.

- The agent's credence fixes a price, and she'll take either side of it.

The Finite Additivity Book

P and Q mutually exclusive. Credence 0.2 in each, but 0.8 in $P \cup Q$.

Finite additivity requires 0.4.

The Finite Additivity Book

P and Q mutually exclusive. Credence 0.2 in each, but 0.8 in $P \cup Q$.

Finite additivity requires 0.4.

1. She sells a \$100 bet on P for **\$21**. Her price is \$20, so selling above it is a gain.

The Finite Additivity Book

P and Q mutually exclusive. Credence 0.2 in each, but 0.8 in $P \cup Q$.

Finite additivity requires 0.4.

1 She sells a \$100 bet on P for **\$21**. Her price is \$20, so selling above it is a gain.
2 She sells a \$100 bet on Q for **\$21**. Likewise.

The Finite Additivity Book

P and Q mutually exclusive. Credence 0.2 in each, but 0.8 in $P \cup Q$.

Finite additivity requires 0.4.

1. She sells a \$100 bet on P for **\$21**. Her price is \$20, so selling above it is a gain.
2. She sells a \$100 bet on Q for **\$21**. Likewise.
3. She buys a \$100 bet on $P \cup Q$ for **\$79**. Her price is \$80, so buying below it is a gain.

The Finite Additivity Book

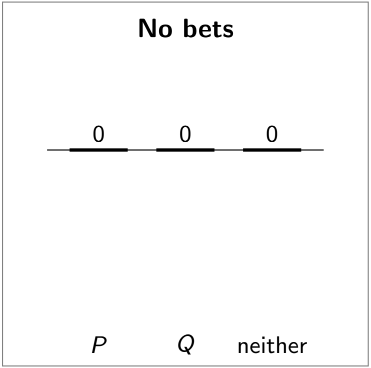
P and Q mutually exclusive. Credence 0.2 in each, but 0.8 in $P \cup Q$.
Finite additivity requires 0.4.

1. She sells a \$100 bet on P for **\$21**. Her price is \$20, so selling above it is a gain.
2. She sells a \$100 bet on Q for **\$21**. Likewise.
3. She buys a \$100 bet on $P \cup Q$ for **\$79**. Her price is \$80, so buying below it is a gain.

Each one, on its own, is one her own credences say she wants.

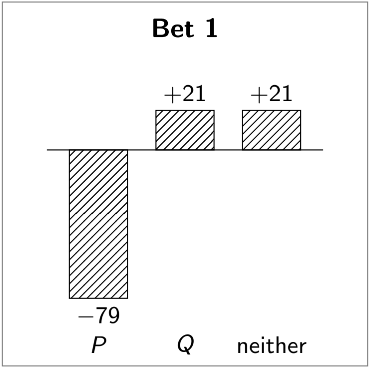
Her Net Position

No bets taken.



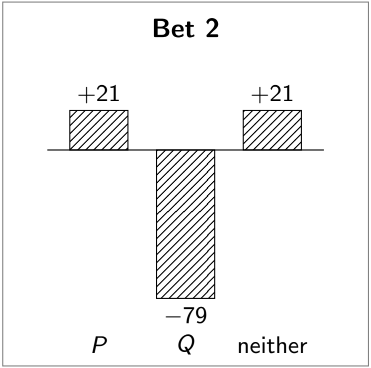
Her Net Position

Bet 1 on its own — she sold a \$100 bet on P for \$21.



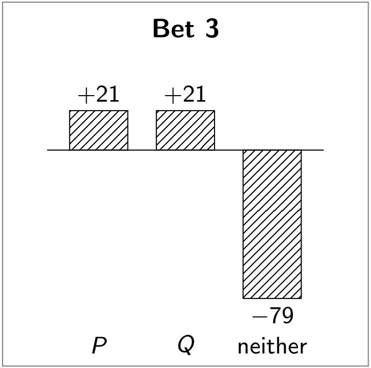
Her Net Position

Bet 2 on its own — she sold a \$100 bet on Q for \$21.



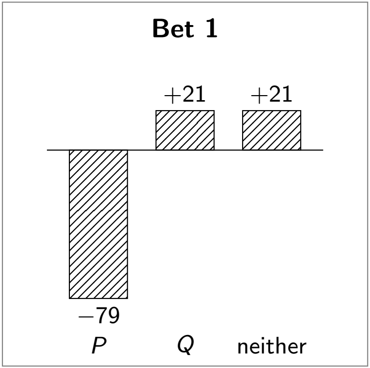
Her Net Position

Bet 3 on its own — she bought a \$100 bet on $P \cup Q$ for \$79.



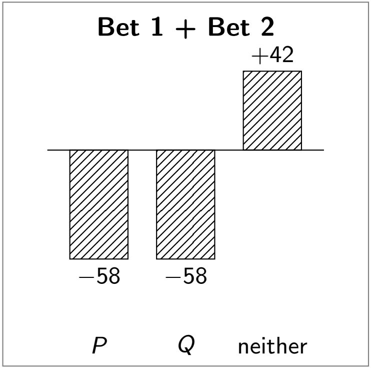
Her Net Position

Now accumulating. After **Bet 1** she is exposed only if P happens.



Her Net Position

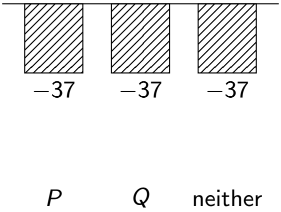
After **Bets 1 and 2** — down in both, but still ahead if neither happens.



Her Net Position

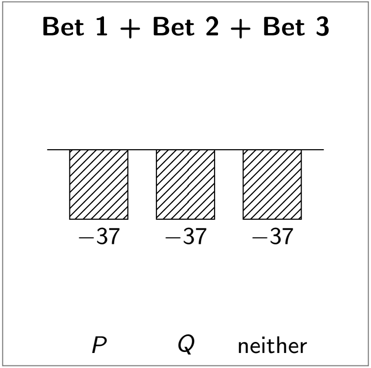
After **all three**.

Bet 1 + Bet 2 + Bet 3



Her Net Position

After **all three**.



She is down $37 whatever happens.

Dutch Books as Arguments for Probabilism

- To get an argument for probabilism out of the Dutch Book we need to explain what's wrong with exploitability.

Dutch Books as Arguments for Probabilism

- To get an argument for probabilism out of the Dutch Book we need to explain what's wrong with exploitability.
- The flat-footed version runs: if your credences are non-probabilistic, a cunning Dutchman might come along and take your money; if they are probabilistic, he cannot; so you had better have probabilistic credences.

Why That Isn't Good Enough

- On this version of the argument, the reason why we should not have non-probabilistic credences is that avoiding them protects us from actually being exploited in a Dutch book situation.

Why That Isn't Good Enough

- On this version of the argument, the reason why we should not have non-probabilistic credences is that avoiding them protects us from actually being exploited in a Dutch book situation.
- However, suppose that for some reason an agent knew for certain that no Dutch book situation would arise. In that case they would know that having non-probabilistic credences will not hurt them, and **the reason to adopt probabilistic credences would vanish**.

Why That Isn't Good Enough

- On this version of the argument, the reason why we should not have non-probabilistic credences is that avoiding them protects us from actually being exploited in a Dutch book situation.
- However, suppose that for some reason an agent knew for certain that no Dutch book situation would arise. In that case they would know that having non-probabilistic credences will not hurt them, and **the reason to adopt probabilistic credences would vanish**.
- But presumably a proponent of probabilism would not want to say that an agent like this is rational: although they will not be exploited, **it is their pattern of credences itself which is supposed to be irrational**.

A Better Version

1. An agent is rationally required to bet in line with the betting interpretation, given the credences they have.

A Better Version

1. An agent is rationally required to bet in line with the betting interpretation, given the credences they have.
2. If an agent has non-probabilistic credences and bets in line with the betting interpretation, then they are vulnerable to a Dutch book.

A Better Version

1. An agent is rationally required to bet in line with the betting interpretation, given the credences they have.
2. If an agent has non-probabilistic credences and bets in line with the betting interpretation, then they are vulnerable to a Dutch book.
3. If an agent is vulnerable to a Dutch book, then their credences are irrational.

A Better Version

1. An agent is rationally required to bet in line with the betting interpretation, given the credences they have.
2. If an agent has non-probabilistic credences and bets in line with the betting interpretation, then they are vulnerable to a Dutch book.
3. If an agent is vulnerable to a Dutch book, then their credences are irrational.

$\therefore$ If an agent has non-probabilistic credences, then their credences are irrational.

The Betting Interpretation

- You're obviously not required to maximise expected *money*.

The Betting Interpretation

- You're obviously not required to maximise expected *money*.
- So you're not in general required to bet in line with the betting interpretation.

The Betting Interpretation

- You're obviously not required to maximise expected *money*.
- So you're not in general required to bet in line with the betting interpretation.
- There's an answer: the premise is far more plausible when the stakes are **small**.

The Betting Interpretation

- You're obviously not required to maximise expected *money*.
- So you're not in general required to bet in line with the betting interpretation.

- There's an answer: the premise is far more plausible when the stakes are **small**.
- And however small they're made, you can still be made to lose for sure.

Two Boxes

- A transparent box, holding \$1,000. You can see that it does.

Two Boxes

- A transparent box, holding \$1,000. You can see that it does.
- A blue box, holding either \$1,000,000 or nothing.

Two Boxes

- A transparent box, holding \$1,000. You can see that it does.
- A blue box, holding either \$1,000,000 or nothing.
- **You may take the blue box alone, or you may take both.**

Newcomb's Problem (with added demons)

- Last night, while you slept, a demon scanned your brain and predicted what you would do in *Two Boxes*.

Newcomb's Problem (with added demons)

- Last night, while you slept, a demon scanned your brain and predicted what you would do in *Two Boxes*.
- If the demon predicted you would take one box: they put \$1,000,000 in the blue box.

Newcomb's Problem (with added demons)

- Last night, while you slept, a demon scanned your brain and predicted what you would do in *Two Boxes*.
- If the demon predicted you would take one box: they put $1,000,000 in the blue box.
- If the demon predicted you would take both boxes: they left the blue box empty.

Newcomb's Problem (with added demons)

- Last night, while you slept, a demon scanned your brain and predicted what you would do in *Two Boxes*.
- If the demon predicted you would take one box: they put $1,000,000 in the blue box.
- If the demon predicted you would take both boxes: they left the blue box empty.
- The demon has done this many times, and has almost always been right.

The Case For Two-Boxing

The money is already in the blue box, or it's already not there.

The Case For Two-Boxing

The money is already in the blue box, or it's already not there.

Nothing you do now changes it.

So - take the extra money!

The Case for One-Boxing

People who take one box almost always end up millionaires. People who take both boxes almost always end up with only one thousand dollars.

The Case for One-Boxing

People who take one box almost always end up millionaires. People who take both boxes almost always end up with only one thousand dollars.

If you take only one box, you are way more likely to be one of the millionaires.

So - take one box and enjoy being rich!

What Would You Do?

One box, or two boxes?

The Obvious Table

Taking the states of nature to be what the demon predicted last night.

	Predicted one box	Predicted both
Take the blue box	\$1,000,000	\$0
Take both boxes	\$1,001,000	\$1,000

The Obvious Table

Taking the states of nature to be what the demon predicted last night.

	Predicted one box	Predicted both
Take the blue box	\$1,000,000	\$0
Take both boxes	\$1,001,000	\$1,000

- Taking both gets you \$1,000 more **in either column**.

The Obvious Table

Taking the states of nature to be what the demon predicted last night.

	Predicted one box	Predicted both
Take the blue box	\$1,000,000	\$0
Take both boxes	\$1,001,000	\$1,000

- Taking both gets you \$1,000 more **in either column**.
- So taking both **strongly statewise dominates**.

Are These the Right States of Nature?

Recall the conditions we had before on states of nature:

Are These the Right States of Nature?

Recall the conditions we had before on states of nature:

(i) Together with the act, they resolve the uncertainty and guarantee one outcome.

(ii) The act you choose does not affect whether the state occurs, or how probable it is.

(iii) They are mutually exclusive.

(iv) They are jointly exhaustive.

Should States of Nature Satisfy Probabilistic Independence?

- (i), (iii) and (iv) are all fine.

Should States of Nature Satisfy Probabilistic Independence?

- (i), (iii) and (iv) are all fine.
- But your choice is **strong evidence** about what the demon predicted.

Should States of Nature Satisfy Probabilistic Independence?

- (i), (iii) and (iv) are all fine.
- But your choice is **strong evidence** about what the demon predicted.

Write B for taking both boxes, D for the demon predicting that you would take both boxes:

$$\mathrm{Cr}(D \mid B) \approx 1, \qquad \mathrm{Cr}(D \mid B^c) \approx 0.$$

$$\text{So} \quad \mathrm{Cr}(D \mid B) \neq \mathrm{Cr}(D).$$

Should States of Nature Satisfy Probabilistic Independence?

- (i), (iii) and (iv) are all fine.
- But your choice is **strong evidence** about what the demon predicted.

Write B for taking both boxes, D for the demon predicting that you would take both boxes:

$$\mathrm{Cr}(D \mid B) \approx 1, \qquad \mathrm{Cr}(D \mid B^c) \approx 0.$$

$$\text{So} \quad \mathrm{Cr}(D \mid B) \neq \mathrm{Cr}(D).$$

Which is exactly the failure of probabilistic independence.

What That Shows, and What It Doesn't

- It does **not** prove the states are wrong.

What That Shows, and What It Doesn't

- It does **not** prove the states are wrong.
- It shows there's a **dispute** about what the relevant states should be.

What That Shows, and What It Doesn't

- It does **not** prove the states are wrong.
- It shows there's a **dispute** about what the relevant states should be.
- Condition (ii) can be read as requiring two things that are incompatible.

What That Shows, and What It Doesn't

- It does **not** prove the states are wrong.
- It shows there's a **dispute** about what the relevant states should be.
- Condition (ii) can be read as requiring two things that are incompatible.
  (a) **Probabilistic independence.**

What That Shows, and What It Doesn't

- It does **not** prove the states are wrong.
- It shows there's a **dispute** about what the relevant states should be.

- Condition (ii) can be read as requiring two things that are incompatible.
  (a) **Probabilistic independence.**
  (b) **Causal independence.**

Probabilistic and Causal Independence

Probabilistic independence

The set of states of nature is probabilistically independent of the acts if and only if, for every state S_i and every act A,

$$\mathrm{Cr}(S_i \mid A) = \mathrm{Cr}(S_i).$$

Probabilistic and Causal Independence

Probabilistic independence

The set of states of nature is probabilistically independent of the acts if and only if, for every state S_i and every act A,

$$Cr(S_i \mid A) = Cr(S_i).$$

Causal independence

The set of states of nature is causally independent of the acts if and only if which action we take **makes** no difference to the probabilities of the states.

Probabilistic and Causal Independence

Probabilistic independence

The set of states of nature is probabilistically independent of the acts if and only if, for every state S_i and every act A,

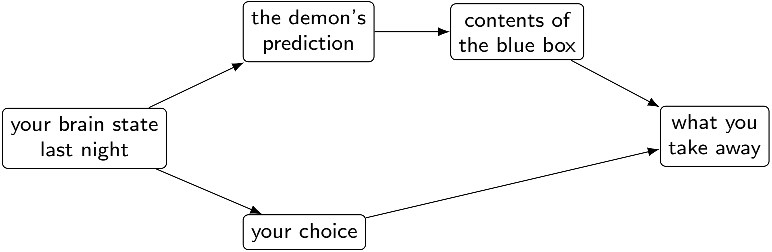
$$\mathrm{Cr}(S_i \mid A) = \mathrm{Cr}(S_i).$$

Causal independence

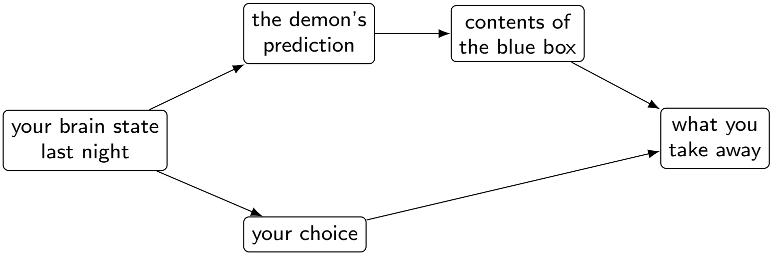
The set of states of nature is causally independent of the acts if and only if which action we take **makes** no difference to the probabilities of the states.

- This second notion isn't as clearly defined as the first. But it is intuitive.

Why They Came Apart

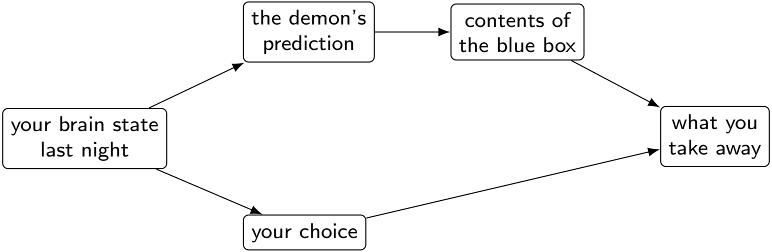


Why They Came Apart



- A **common cause**, and **no arrow** from your choice to the prediction.

Why They Came Apart



- A **common cause**, and **no arrow** from your choice to the prediction.
- Which is why the correlation isn't an influence.

Two State Conditions, Two Theories

- Require **probabilistic independence** of the acts:
  $\rightarrow$ *evidential decision theory*.

Two State Conditions, Two Theories

- Require **probabilistic independence** of the acts:
  $\rightarrow$ *evidential decision theory*.
- Require **causal independence** of the acts:
  $\rightarrow$ *causal decision theory*.

Evidential Decision Theory: Formulation 1

- Read the requirement as **probabilistic** independence.

Evidential Decision Theory: Formulation 1

- Read the requirement as **probabilistic** independence.
- Then each state has one credence, whichever act we are evaluating.

Evidential Decision Theory: Formulation 1

- Read the requirement as **probabilistic** independence.
- Then each state has one credence, whichever act we are evaluating.

(Evidential) expected utility

$$\mathrm{EU}_e(A) = \sum_{i=1}^{n} \mathsf{Cr}(S_i)\, U(A, S_i)$$

Evidential Decision Theory: Formulation 1

- Read the requirement as **probabilistic** independence.
- Then each state has one credence, whichever act we are evaluating.

(Evidential) expected utility

$$\mathrm{EU}_e(A) = \sum_{i=1}^{n} \mathsf{Cr}(S_i)\, U(A, S_i)$$

- **Both theories accept this equation.** They disagree about how the states are to be construed. This is evidential decision theory only if the states are **probabilistically** independent of your actions.

Evidential Decision Theory: Formulation 2

- With a second formula, we don't actually need the states to be probabilistically independent of the acts.

Evidential Decision Theory: Formulation 2

- With a second formula, we don't actually need the states to be probabilistically independent of the acts.

Evidential expected utility, generalised

$$\mathrm{EU}_e(A) = \sum_{i=1}^{n} \mathsf{Cr}(S_i \mid A)\, U(A, S_i)$$

Evidential Decision Theory: Formulation 2

- With a second formula, we don't actually need the states to be probabilistically independent of the acts.

Evidential expected utility, generalised

$$\mathrm{EU}_e(A) = \sum_{i=1}^{n} \mathsf{Cr}(S_i \mid A)\, U(A, S_i)$$

- All we now require is that each state, with each act, settles everything you care about.

Evidential Decision Theory: Formulation 2

- With a second formula, we don't actually need the states to be probabilistically independent of the acts.

Evidential expected utility, generalised

$$\mathrm{EU}_e(A) = \sum_{i=1}^{n} \mathsf{Cr}(S_i \mid A)\, U(A, S_i)$$

- All we now require is that each state, with each act, settles everything you care about.
- And where the states *are* independent, $\mathsf{Cr}(S_i \mid A) = \mathsf{Cr}(S_i)$ — so the two agree.

Evidential Decision Theory: Formulation 2

- With a second formula, we don't actually need the states to be probabilistically independent of the acts.

Evidential expected utility, generalised

$$\mathrm{EU}_e(A) = \sum_{i=1}^{n} \mathsf{Cr}(S_i \mid A)\, U(A, S_i)$$

- All we now require is that each state, with each act, settles everything you care about.
- And where the states *are* independent, $\mathsf{Cr}(S_i \mid A) = \mathsf{Cr}(S_i)$ — so the two agree.
- Notice that dominance reasoning does not apply for such partitions.

Newcomb, on the Evidential Theory

Taking the demon to be right 95% of the time.

Newcomb, on the Evidential Theory

Taking the demon to be right 95% of the time.

$$
\begin{aligned}
\mathrm{EU}_e(\text{one box}) &= 1{,}000{,}000(0.95) + 0(0.05) &&= 950{,}000 \\
\mathrm{EU}_e(\text{both}) &= 1{,}001{,}000(0.05) + 1{,}000(0.95) &&= 51{,}000
\end{aligned}
$$

So evidential decision theory says take one box.

Newcomb, on the Evidential Theory

Taking the demon to be right 95% of the time.

$$
\begin{aligned}
\mathrm{EU}_e(\text{one box}) &= 1{,}000{,}000(0.95) + 0(0.05) &&= 950{,}000 \\
\mathrm{EU}_e(\text{both}) &= 1{,}001{,}000(0.05) + 1{,}000(0.95) &&= 51{,}000
\end{aligned}
$$

Newcomb, on the Evidential Theory

Taking the demon to be right 95% of the time.

$$\begin{aligned} \mathrm{EU}_e(\text{one box}) &= 1{,}000{,}000(0.95) + 0(0.05) &&= 950{,}000 \\ \mathrm{EU}_e(\text{both}) &= 1{,}001{,}000(0.05) + 1{,}000(0.95) &&= 51{,}000 \end{aligned}$$

So evidential decision theory says take one box.

Causal Decision Theory

- Take the other reading: the states must be **causally** independent of what you do.

Causal Decision Theory

- Take the other reading: the states must be **causally** independent of what you do.
- All the work now goes into saying which states those are.

Causal Decision Theory

- Take the other reading: the states must be **causally** independent of what you do.
- All the work now goes into saying which states those are.

Dependency hypothesis (Lewis)

A maximally specific proposition about how the things you care about do and do not depend causally on your present actions.

What One Actually Looks Like

- Take one counterfactual for **each** option: *if I were to do A, the outcome would be O*.

What One Actually Looks Like

- Take one counterfactual for **each** option: *if I were to do A, the outcome would be O*.
- A dependency hypothesis is a whole **pattern** of these — one per option, all at once.

What One Actually Looks Like

- Take one counterfactual for **each** option: *if I were to do A, the outcome would be O*.
- A dependency hypothesis is a whole **pattern** of these — one per option, all at once.

Lewis's example. Three options, and Bruce the cat might purr loudly, softly, or not at all.

What One Actually Looks Like

- Take one counterfactual for **each** option: *if I were to do A, the outcome would be O*.
- A dependency hypothesis is a whole **pattern** of these — one per option, all at once.

Lewis's example. Three options, and Bruce the cat might purr loudly, softly, or not at all.

I brush Bruce $\Box\!\!\rightarrow$ he purrs loudly
I stroke Bruce $\Box\!\!\rightarrow$ he purrs softly
I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr

Newcomb's Dependency Hypotheses

K_1: one box $\Box\!\!\rightarrow$ \$1,000,000
both boxes $\Box\!\!\rightarrow$ \$1,001,000

Newcomb's Dependency Hypotheses

K_1: one box $\Box\!\!\rightarrow$ \$1,000,000
both boxes $\Box\!\!\rightarrow$ \$1,001,000

K_2: one box $\Box\!\!\rightarrow$ \$0
both boxes $\Box\!\!\rightarrow$ \$1,000

Newcomb's Dependency Hypotheses

$$K_1: \quad \text{one box} \;\Box\!\!\rightarrow \$1{,}000{,}000$$
$$\text{both boxes} \;\Box\!\!\rightarrow \$1{,}001{,}000$$

$$K_2: \quad \text{one box} \;\Box\!\!\rightarrow \$0$$
$$\text{both boxes} \;\Box\!\!\rightarrow \$1{,}000$$

- Neither of these says what you *will* choose — only what your choosing affects.

Utility

Causal expected utility

$$\mathrm{EU}_c(A) = \sum_K \mathsf{Cr}(K)\, U(A,K)$$

Utility

Causal expected utility

$$\mathrm{EU}_c(A) = \sum_K \mathsf{Cr}(K)\, U(A,K)$$

- Note that the credences are unconditional.

Utility

Causal expected utility

$$\mathrm{EU}_c(A) = \sum_K \mathrm{Cr}(K)\, U(A, K)$$

- Note that the credences are unconditional.
- Your act might be evidence about which K holds — just as it was about the prediction.

Utility

Causal expected utility

$$\mathrm{EU}_c(A) = \sum_K \mathsf{Cr}(K)\, U(A, K)$$

- Note that the credences are unconditional.
- Your act might be evidence about which K holds — just as it was about the prediction.
- Conditioning the credence in K on act A would let that evidence back in.

Causey and the LAN Outlet

	Programme A (LAN predicted touch)	Programme B (LAN predicted no touch)
no touch	worse shock	nothing
touch	regular shock	nice melody

Evidentia and the Prediction Unit

	DPU predicted shock	DPU predicted no shock
shock today	1 shock total	~~21 shocks total~~
no shock today	~~0 shock total~~	20 shocks total

Evidentia and the Prediction Unit

	DPU predicted shock	DPU predicted no shock
shock today	1 shock total	~~21 shocks total~~
no shock today	~~0 shock total~~	20 shocks total

The crossed-out cells are the ones she is almost certain do not obtain.

Why Ain'cha Rich?

- One-boxers walk away with a million. Two-boxers with a thousand. Over and over.

Why Ain'cha Rich?

- One-boxers walk away with a million. Two-boxers with a thousand. Over and over.
- **Doesn't that settle it?**

Why Ain'cha Rich?

- One-boxers walk away with a million. Two-boxers with a thousand. Over and over.
- **Doesn't that settle it?**
- Here's a counterargument on the part of the CDT-ist

Why Ain'cha Rich?

- One-boxers walk away with a million. Two-boxers with a thousand. Over and over.
- **Doesn't that settle it?**

- Here's a counterargument on the part of the CDT-ist
- People who apply EDT are being rewarded by the demon BEFORE they take their decision. By CDT lights, they're being rewarded for being irrational.

Why Ain'cha Rich?

- One-boxers walk away with a million. Two-boxers with a thousand. Over and over.
- **Doesn't that settle it?**

- Here's a counterargument on the part of the CDT-ist
- People who apply EDT are being rewarded by the demon BEFORE they take their decision. By CDT lights, they're being rewarded for being irrational.
- But that doesn't prove that they're rational.

Why Ain'cha Rich?

- One-boxers walk away with a million. Two-boxers with a thousand. Over and over.
- **Doesn't that settle it?**

- Here's a counterargument on the part of the CDT-ist
- People who apply EDT are being rewarded by the demon BEFORE they take their decision. By CDT lights, they're being rewarded for being irrational.
- But that doesn't prove that they're rational.
- However, it turns out that the situation is asymmetric: this can happen to the CDT-ist, but not to the EDT-ist.
- If being "irrational" is rewarded, EDT-ists will do the "irrational" thing associated with the reward.

The Lesion

A gene causes both a desire to smoke and cancer. Smoking itself causes nothing bad.

	Cancer	No cancer
Smoke	-9	1
Don't smoke	-10	0

The Lesion

A gene causes both a desire to smoke and cancer. Smoking itself causes nothing bad.

	Cancer	No cancer
Smoke	-9	1
Don't smoke	-10	0

With $\mathrm{Cr}(C \mid S) = 0.8$ and $\mathrm{Cr}(C \mid S^c) = 0.2$:

$$\mathrm{EU}_e(S) = -7, \qquad \mathrm{EU}_e(S^c) = -2.$$

The Lesion

A gene causes both a desire to smoke and cancer. Smoking itself causes nothing bad.

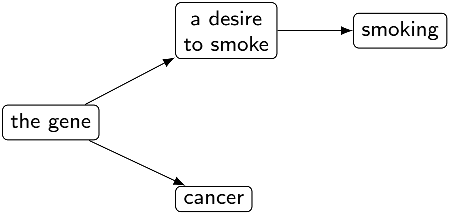
	Cancer	No cancer
Smoke	-9	1
Don't smoke	-10	0

With $\mathrm{Cr}(C \mid S) = 0.8$ and $\mathrm{Cr}(C \mid S^c) = 0.2$:

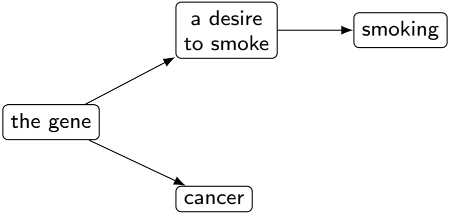
$$\mathrm{EU}_e(S) = -7, \qquad \mathrm{EU}_e(S^c) = -2.$$

So the evidential theory says give up smoking.

Another Causal Diagram

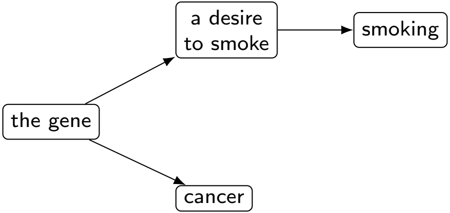


Another Causal Diagram



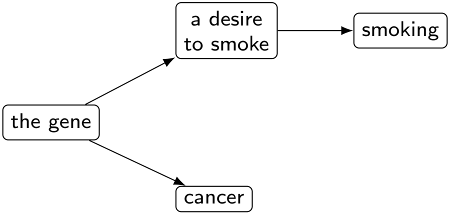
- The gene causes both cancer and a desire to smoke; and the desire to smoke causes people to smoke; but smoking does not cause cancer.

Another Causal Diagram



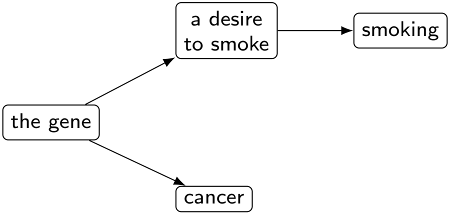
Legal notice and full disclaimer. The foregoing slide is a stipulated philosophical example and is **not** a statement of fact, opinion, medical advice, or public-health guidance. Dr Francis wishes to state, in the clearest possible terms and without any qualification whatsoever, that **smoking does cause lung cancer**, and that nothing said, written, projected, implied, gestured at, or inadvertently conveyed in the course of this lecture, this module, this semester, or the remainder of his natural life should be construed otherwise. Smoking also causes, without limitation: cancers of the mouth, throat, larynx, oesophagus, stomach, pancreas, liver, kidney, bladder, cervix, bowel and blood; chronic obstructive pulmonary disease; emphysema; chronic bronchitis; coronary heart disease; stroke; peripheral vascular disease; abdominal aortic aneurysm; type 2 diabetes; rheumatoid arthritis; osteoporosis; macular degeneration; cataracts; reduced fertility in both sexes; complications of pregnancy including low birth weight, preterm delivery and stillbirth; delayed wound healing; periodontal disease; and premature ageing of the skin. Second-hand smoke causes many of the same conditions in persons who have not smoked at all, including children, who are not in a position to consent to it. The example on this slide posits a fictitious gene which is a common cause of a desire

Another Causal Diagram



- The gene causes both cancer and a desire to smoke; and the desire to smoke causes people to smoke; but smoking does not cause cancer.
- **It makes no difference whether smoking causes cancer or merely indicates it.**

Another Causal Diagram



- The gene causes both cancer and a desire to smoke; and the desire to smoke causes people to smoke; but smoking does not cause cancer.
- **It makes no difference whether smoking causes cancer or merely indicates it.**
- Either way, smoking is bad news.

The Tickle Defence

- The gene causes smoking by causing the desire.

The Tickle Defence

- The gene causes smoking by causing the desire.
- If you already know whether you have the desire, the desire is already the evidence.

The Tickle Defence

- The gene causes smoking by causing the desire.
- If you already know whether you have the desire, the desire is already the evidence.
- Acting on it tells you nothing further.

The Tickle Defence

- The gene causes smoking by causing the desire.
- If you already know whether you have the desire, the desire is already the evidence.
- Acting on it tells you nothing further.
- So $\mathrm{Cr}(C \mid S) = \mathrm{Cr}(C \mid S^c)$, and the case doesn't arise.

Death in Damascus

Tomi meets Death in Damascus, flees to Aleppo, and finds Death waiting. The appointment book said Aleppo all along.

	Death seeks Aleppo	Death seeks Damascus
Go to Aleppo	dies (-100)	lives (0)
Stay in Damascus	lives (0)	dies (-100)

Death in Damascus

Tomi meets Death in Damascus, flees to Aleppo, and finds Death waiting. The appointment book said Aleppo all along.

	Death seeks Aleppo	Death seeks Damascus
Go to Aleppo	dies (-100)	lives (0)
Stay in Damascus	lives (0)	dies (-100)

With $\mathrm{Cr}(S_A \mid A) = \mathrm{Cr}(S_D \mid D) = 0.99$, and $a = \mathrm{Cr}(A)$:

$$U(A) = -1 - 98a, \qquad U(D) = -99 + 98a.$$

Unstable

- $U(D) > U(A)$ whenever $a > 1/2$.

Unstable

- $U(D) > U(A)$ whenever $a > 1/2$.
- $U(A) > U(D)$ whenever $a < 1/2$.

Unstable

- $U(D) > U(A)$ whenever $a > 1/2$.
- $U(A) > U(D)$ whenever $a < 1/2$.
- Whichever act Tomi thinks he will perform, the other one looks better.

Unstable

- $U(D) > U(A)$ whenever $a > 1/2$.
- $U(A) > U(D)$ whenever $a < 1/2$.
- Whichever act Tomi thinks he will perform, the other one looks better.
- The decision is unstable in the sense that making a decision to perform act A seems to make performing act A no longer rational.

Unstable

- $U(D) > U(A)$ whenever $a > 1/2$.
- $U(A) > U(D)$ whenever $a < 1/2$.
- Whichever act Tomi thinks he will perform, the other one looks better.
- The decision is unstable in the sense that making a decision to perform act A seems to make performing act A no longer rational.
- The paradox: it seems impossible to rationally decide and follow through. But it must be rational to decide and follow through on *something*!

Next Week

Expected utility theory

- BDRC ch 6: core
- BDRC ch 5: recommended but optional