Counterfactual reasoning: Lecture 1

Tsinghua Logic Summer School

Tadeg Quillien

University of Edinburgh

2026

The ladder of causality

In the first lecture we said that there were three different kinds of queries we can ask from a causal model:

- if we see that $X = x$, then what is the value of Y?
- if we Do($X = x$), what is the value of Y?
- If X had been x (instead of the value it currently has), what would have been the value of Y?

So far we have seen how to compute the first two kinds of queries.

- We now turn to the third kind: **counterfactual** queries.

Notation

The value of Y, if X had been x, is denoted as $Y_{X=x}$

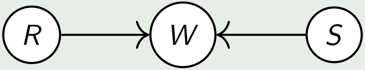
Uses of counterfactual reasoning

- Guide to future decision-making.
  - ‘I arrived home early (but I almost crashed into a car).’
  - ‘I lost the game, but if I had played more defensively I would have won.’
- Attribution of responsibility
  - ‘If you had not forgotten to water the plants they wouldn’t have died.’
- Explanation
  - ‘If you had a better credit score your loan would have been approved.’

Example of counterfactual query

Example

Both the Rain and the Sprinkler make the pavement Wet:



$W := R \vee S$.

- Today, it Rained, the Sprinkler was on, and the Pavement was Wet ($R = 1$, $S = 1$, $W = 1$).
- If the Sprinkler had been off, would the Pavement still be Wet?

Intuitively: **yes**. How do we formalize this?

The intuition

Example

$$R \longrightarrow W \longleftarrow S$$

$W := R \vee S$.

- Today, it Rained, the Sprinkler was on, and the Pavement was Wet ($R = 1,\ S = 1,\ W = 1$).
- If the Sprinkler had been off, would the Pavement still be Wet?

We **intervene** to turn off the Sprinkler ($\mathrm{Do}(S = 0)$) **while keeping everything else the same**.

- We pass our new value for S to the equation for W:
- $W := R \vee S = 1 \vee 0$

$\rightarrow$ The Pavement is still Wet!

Counterfactual inference

Definition

In an SCM, a counterfactual inference of the form 'If X had been x, would Y had been y?', is computed as:

$$P(Y = y_{X=x} | \mathcal{U} = \vec{u}) = P(Y = y | \mathrm{Do}(X = x), \mathcal{U} = \vec{u})$$

- ($Y = y_{X=x}$ is new notation that denotes a counterfactual query.)

Remark

- Conditioning on $\mathcal{U} = \vec{u}$ means that we 'start' from what happened in the actual world.
- Because SCMs are deterministic, $\mathcal{U} = \vec{u}$ contains all the information we need about what happened.

Counterfactual inference

Example

$$R \longrightarrow W \longleftarrow S$$

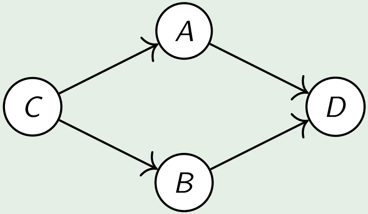
$W := R \vee S$. In actual world, $R = 1$, $S = 1$, $W = 1$.

$$\begin{aligned} & P(W = 1_{R=0} | \mathcal{U} = \vec{u}) \\ = & P(W = 1 | \mathrm{Do}(R = 0), R = 1, S = 1) \\ = & 1 \end{aligned}$$

Making things a bit more complicated

Example

If the Captain (C) gives the order, soldiers A and B shoot. If either soldier shoots, the prisoner Dies (D).



We know that soldier A shot and that the prisoner Died, but we don't know the state of C and B.

If A had not shot, would the prisoner still have died?

Counterfactual reasoning

Counterfactual reasoning follows a three-steps recipe: (informal)

- Infer unobserved events,
- Make an intervention,
- Compute the effect of the intervention.

Example

- **Infer** that the captain gave the order,
- **Intervene** on soldier A,
- **Compute** that soldier B shoots anyway, so $D = 1$.

Note : sometimes the inference is probabilistic.

- e.g. if soldiers sometimes shoot on their own.

Counterfactual reasoning with partial information

Reminder

$$P(Y = y_{X=x} | \mathcal{U} = \vec{u}) = P(Y = y | \mathrm{Do}(X = x), \mathcal{U} = \vec{u})$$

Suppose we observe only some variables $\mathcal{Z}$. Then the counterfactual query $Y = y_{X=x}$ can be computed as:

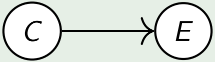
$$P(Y = y_{X=x} | \mathcal{Z} = \vec{z})$$
$$= \sum_{\vec{u}} P(Y = y | \mathrm{Do}(X = x), \mathcal{U} = \vec{u}) P(\mathcal{U} = \vec{u} | \mathcal{Z} = \vec{z}).$$

Explanation

- $P(\mathcal{U} = \vec{u} | \mathcal{Z} = \vec{z})$: Infer what happened.
- $P(Y = y | \mathrm{Do}(X = x)$ Make a prediction assuming $\mathcal{U} = \vec{u}$.
- $\sum_{\vec{u}}$: sum over possible states of $\mathcal{U}$.

Counterfactual reasoning with partial information

Example



With $E := C \vee U_e$. Suppose $C = 1$, $E = 1$, in the actual world.
We have:

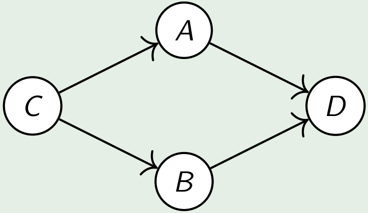
$$P(E_{\neg C} | C = 1, E = 1) = \sum_{U_e} P(E|\text{Do}(\neg C), Ue) P(U_e | C, E).$$

We can easily compute that $P(Ue|C, E) = P(U_e)$. So the equation above becomes:

$$\begin{aligned} P(E|\text{Do}(\neg C), U_e)P(U_e) + P(E|\text{Do}(\neg C), \neg U_e)P(\neg U_e) \\ = 1 \times P(U_e) + 0 \times P(\neg U_e) = P(U_e). \end{aligned}$$

Difference between Level-2 and Level-3 queries

Example



Suppose that $P(C) = .1$. Then:

- The **Level-2** query $P(D = 1|\mathrm{Do}(A = 0))$ returns a low value (.1),
- While the **Level-3** query $P(D = 1_{A=0}|A = 1, D = 1)$ returns value 1.

Why counterfactual reasoning sits on its own rung of the ladder of causality

Counterfactual reasoning requires that we infer the value of all variables in $\mathcal{U}$.

But in going from an SCM to a CBN, some variables in $\mathcal{U}$ get 'dissolved' in the conditional distributions. E.g.:

- SCM: $E := C \wedge U_{EC}$, $P(U_{EC}) = .7$.
- CBN: $P(E|C) = .7$.

$\rightarrow$ Counterfactual reasoning requires SCMs.

Why counterfactual reasoning sits on its own rung on the ladder of causality

Example

Consider the following SCM: $E := A \text{ XOR } U_a$,
With $P(U_a) = .9$.
CBN representation:

- $P(E|A) = .1$, $P(E|\neg A) = .9$

In the actual world we have $A = 1$, $E = 1$. What would have happened if A had been 0?
With SCM:

- We infer that $U_a = 0$. $\rightarrow$ So E would have been 0.

With CBN:

- We can't infer U_a because it isn't represented in the CBN.
- We know that $P(E|\text{Do}(\neg A)) = .9$, but this is not what we want (and doesn't give the right answer).

Another way to understand the problem with CBNs

Example

Remember the CBN from last slide:

- $P(E|A) = .1$, $P(E|\neg A) = .9$, with $A = 1$, $E = 1$ in the actual world.

It is consistent with several different SCMs:

- $\mathcal{M}_1$: $E := A \text{ XOR } U_a$, $P(U_a) = .9$
- $\mathcal{M}_2$: $E := (\neg A \wedge U_{AE}) \vee U_a$, $P(U_{AE}) = 8/9$, $P(U_a) = .1$.

Each SCM makes a different counterfactual prediction!

- $\mathcal{M}_1$: If $A = 0$, E would be 0.
- $\mathcal{M}_2$: If $A = 0$, E would be 1.

Observational counterfactual reasoning

Example

- 'What if we had forced soldier A not to shoot?' $\rightarrow$ Interventional counterfactual (this lecture),
- 'What if we had observed soldier A not to shoot?' $\rightarrow$ Observational counterfactual.

Intuitively, if we had observed soldier A not shoot, the prisoner would not have died.
The theory of counterfactuals in this lecture cannot deal with observational counterfactuals.

- But see Lucas & Kemp (2015) if you are curious to learn more.

Remark

Despite being called 'observational', these counterfactuals very much belong to Level 3 on the ladder of causality.

Summary

The ladder of causation:

Level 1: Observational queries. Can be done with Bayes Nets, CBNs, or SCMs.

Level 2: Interventional queries. Can be done with CBNs or SCMs.

Level 3: Counterfactual queries. Requires SCMs.