Counterfactual reasoning: Lecture 2

Tsinghua Logic Summer School

Tadeg Quillien

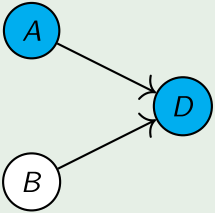
University of Edinburgh

2026

Actual causation

Actual causation

Example



Soldier A shoots, Soldier B doesn't shoot, prisoner Dies.
The fact that A shot caused the prisoner to die. We might want to justify this intuition with counterfactual reasoning:

- If A had not shot, the prisoner would not have died.

Actual causation

Judging actual causation consists in evaluating whether an event caused another one.

Formal statement

Actual causation. Given:

- A Structural Causal Model $\mathcal{M}$,
- A set of actual-world values for all variables in $\mathcal{M}$,

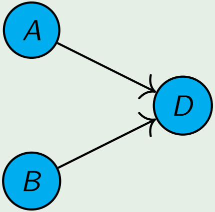
Is event $C = c$ a cause of event $E = e$?

Remark

There isn't really a 'correct' answer.
We are trying to build a formal theory that agrees with our intuitions.

A challenge to the simple counterfactual account

Example



Both soldiers shoot and the prisoner dies.

- If A had not shot, the prisoner would still have died.
- But it still feels like A is responsible for the prisoner's death.

Can we find a more sophisticated theory that involves counterfactual reasoning but delivers the correct verdict?

Halpern's theory (2015)

The main insight is that a cause can be a single event or a **conjunction of events**.

So he would say that $\{A = 1, B = 1\}$ is a cause of $D = 1$.

- Because $D_{A=0,B=0} = 0$.

$A = 1$ is 'part of a cause' because it is a part of $\{A = 1, B = 1\}$.

Other components of Halpern's theory

- **Actuality** requirement: If $C = c$ is a cause $E = e$, then in the actual world $C = c$ and $E = e$.
- **Minimality** requirement: If $\{C\}$ is a cause of E, then $\{C, X\}$ is not a cause of E.
  - e.g. 'Soldier A shot, and the sky is blue' is not a cause of the prisoner dying.

incomplete sketch of Halpern's theory

$\vec{C} = \vec{c}$ is a cause of $E = e$ in model $\mathcal{M}$, with actual world $\mathcal{U} = \vec{u}$ if the following conditions are satisfied:

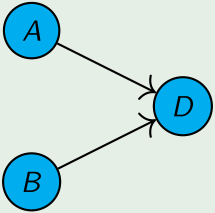
- **Actuality.** In the actual world, $\vec{C} = \vec{c}$ and $E = e$.
- **Counterfactual dependence.** You can perform a counterfactual intervention on $\vec{C}$ that changes the value of E. I.e. there exists a set of values $\vec{c'} \neq \vec{c}$ such that $(E_{\vec{C}=\vec{c'}} | \mathcal{U} = \vec{u}) \neq e$.
- **Minimality**. No subset of $\vec{C} = \vec{c}$ satisfy the above criteria.

Remark

This is an incomplete sketch. We will modify the **counterfactual dependence** criterion later.

Halpern's theory

Example



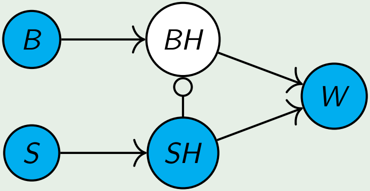
Is $\{A = 1, B = 1\}$ a cause of $D = 1$?

- **Actuality**. In the actual world, $A = 1$, $B = 1$, $E = 1$. ✓
- **Counterfactual dependence**. The counterfactual intervention $\mathrm{Do}(A = 0, B = 0)$ saves the prisoner. ✓
- **Minimality**. Neither $A = 1$ or $B = 1$ on its own satisfies Counterfactual Dependence. ✓

Problem with counterfactual dependence

Example

Billy and Suzy throw a rock at a window. Both throws are perfectly accurate, but Suzy's throw gets there first. The window shatters.
$SH := S$, $BH := B \land \neg SH$, $W := SH \lor BH$.

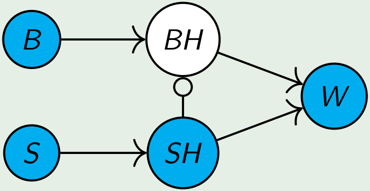


Suzy caused the window to break. But if Suzy had not thrown the rock, the window still would have broken (because of Billy's throw).

Halpern's solution

Example

Billy and Suzy throw a rock at a window. Both throws are perfectly accurate, but Suzy's throw gets there first. The window shatters.
$SH := S$, $BH := B \wedge \neg SH$, $W := SH \vee BH$.



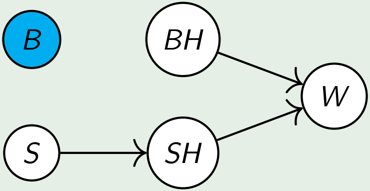
If we hold constant the fact that $BH = 0$:

- Then an intervention preventing Suzy from throwing prevents the window from breaking.

Halpern's solution

Example

Billy and Suzy throw a rock at a window. Both throws are perfectly accurate, but Suzy's throw gets there first. The window shatters.
$SH := S$, $BH := B \wedge \neg SH$, $W := SH \vee BH$.



If we hold constant the fact that $BH = 0$:

- Then an intervention preventing Suzy from throwing prevents the window from breaking.

Halpern's theory (2015)

$\vec{C} = \vec{c}$ is a cause of $E = e$ in model $\mathcal{M}$, with actual world $\mathcal{U} = \vec{u}$ if the following conditions are satisfied:

- **Actuality.** In the actual world, $\vec{C} = \vec{c}$ and $E = e$.
- **Counterfactual dependence.** There exists a set $\vec{W}$ of variables with actual-world values $\vec{w}$ such that, if you Do($\vec{W} = \vec{w}$), you can perform a counterfactual intervention on $\vec{C}$ that changes the value of E. I.e. there exists $\vec{W}$ and $\vec{c'} \neq \vec{c}$ such that $E_{\vec{C}=\vec{c'},\vec{W}=\vec{w}}|\mathcal{U} = \vec{u} \neq e$.
- **Minimality**. No subset of $\vec{C} = \vec{c}$ satisfy the above criteria.

Outstanding challenges

The definition is complicated.

- Maybe a bit ad-hoc?

There are also counter-examples to the definition.
$\rightarrow$ The search for a good theory of actual question is still an unsolved problem.

Causal selection

Example

- Billy will graduate if he passes the Math exam AND the History exam.
- The Math exam is notoriously difficult, but the History exam is very easy.
- Billy passes both exams, and graduates.

Passing Math and passing history are both actual causes of Billy graduating.
But we have the intuition that 'Billy passed because he passed the Math exam'.

$\rightarrow$ We have an intuition that there is **quantitative** variation in the causal responsibility of different events.

Modeling causal selection

Typically, we assign a scalar value (an 'actual causal strength') to each actual cause of an event.

Example

Why did Billy graduate?

- actual causal strength of Math: 0.9.
- actual causal strength of History: 0.2.

Theories of causal selection are outside the scope of this course, but see for example Quillien & Lucas (2023).