Computational Cognitive Science

Lecture 12: Causality 2

Benjamin Peters

School of Informatics

University of Edinburgh

October 23, 2025

slide credits: Chris Lucas

Causality

Optional:

- Chapter 1. Pearl, J. *Causality*. (2009).
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). "A theory of causal learning in children: causal maps and Bayes nets". *Psychological review*.

Causality

Last time, we focused on an associative model of causal learning.

Association can be a clue that *some* causal relationship is at work, but it's neither necessary nor sufficient. Consider:

- The relationship between the gas pedal and car speed when someone is maintaining a constant speed over hills.
- There has been a negative correlation between number of pirates and mean global temperature since the early twentieth century.

Counterfactual theories of causality

"We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects—some of them, at least, and usually all—would have been absent as well" (David Lewis)

- A cause is something that makes a difference
- If the cause had been changed, its effects would have changed

Counterfactual theories of causality

What does it mean to change a cause?

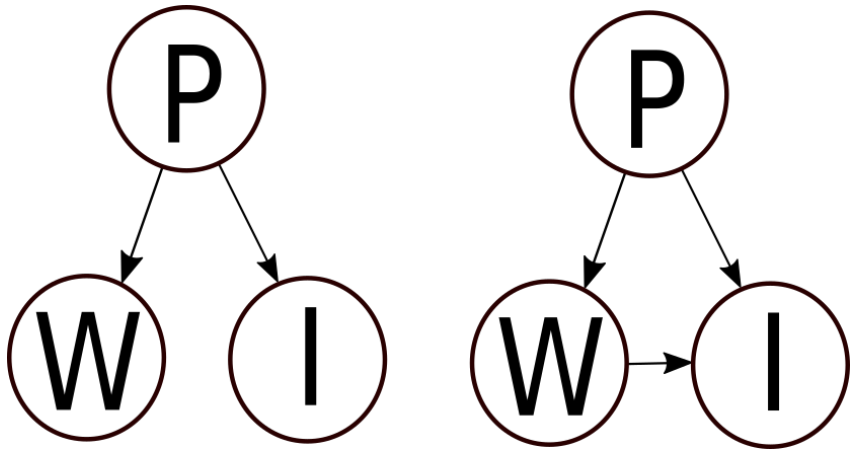
> *"If W were the case, I would be the case" is true in the actual world if and only if either (i) there are no possible W-worlds; or (ii) some W-world where I holds is closer to the actual world than is any W-world where I does not hold." (Lewis; changing variable names)*

How do we judge the closeness of two worlds?

Lewis's give a (complicated) account; but a newer **intervention**-based approach has become popular.

Counterfactual theories of causality

Example: Parties, wine, and insomnia. Suppose we get insomnia after partying and wine; all three are correlated.



In these graphs, arrows (edges) between nodes go from causes to their effects.

Q: How might we distinguish between these two explanations?

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). "A theory of causal learning in children: causal maps and Bayes nets". *Psychological review*.

Interventions

A: Intervene on one or more of the variables; **do** something to them.

What if we take Lewis's most-similar W' world to be one where we **intervene** on W to make it W'?

Pearl: Interventions (real and counterfactual) are all we need to reason about causality.

Interventions

Intuition: The causes of I are the things that change I even when we hold other variables constant.

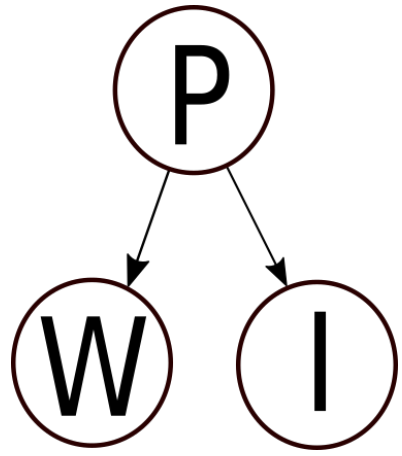
Pearl (2009, p24):

> $P_x(v_i|pa_i) = P(v_i|pa_i)$ *for all* $V_i \notin X$ *whenever* pa_i *is consistent with* $X = x$, *i.e., each* $P(v_i|pa_i)$ *remains invariant to interventions not involving* V_i.

(where pa_i are the *parents* of v_i in a *causal graphical model*)

Causal graphical models

Graphs like this



aren't just convenient visual aids, but part of a powerful causal toolkit.

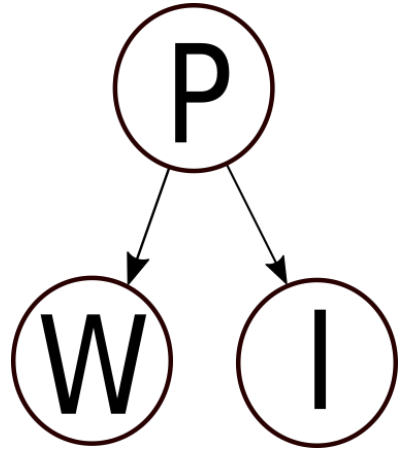
Causal graphical models

Causal graphical models combine

1. Directed graphical models (aka Bayes nets)
2. The *do* operator

Directed graphical models

Graphical models capture probabilistic (in)dependencies. The graph



implies that W and I are marginally dependent, but independent conditional on their shared *parent* P.

Directed graphical models

More generally, a directed graphical model implies that we can factorize the distribution $P(x_1, x_2, ..., x_N)$ into $\prod_{i=1}^{N} P(x_i|pa_i)$ (again, pa_i denotes x_i's parents.)

Bayes nets are **acyclic**, e.g.,

- $A \leftrightarrow B$
- $A \rightarrow B$,$B \rightarrow C$,$C \rightarrow A$

are not valid.

If A causes B and vice versa, we can "unroll" our events in time, e.g., $A_t \rightarrow B_{t+1}$, $B_t \rightarrow A_{t+1}$

Directed graphical models

If we specify Bayes nets corresponding to causal intuitions, they tend to have sensible independences. E.g.,

light switch (s) $\rightarrow$ closed circuit (c) $\rightarrow$ LED illuminates (l)

This factorizes $P(s, c, l)$ into $P(s)P(c|s)P(l|c)$:

- If we condition on a particular value of c, the switch no longer influences the LED
- If we condition on s, c still affects l

Directed graphical models

. . . but this isn't enough. Consider

- bacterial infection $\rightarrow$ infection symptoms
- infection symptoms $\rightarrow$ antibiotics
- bacterial infection $\rightarrow$ death-by-infection $\leftarrow$ antibiotics

If we condition on taking antibiotics, death may be *more* likely.

This is consistent with *observing* someone is taking antibiotics, but doesn't tell us if we should take antibiotics.

Directed graphical models

The joint distribution over switch state, circuit state, and light state $P(s, c, l)$ may be factorized both by[^1]:

- light switch (s) $\rightarrow$ closed circuit (c) $\rightarrow$ LED illuminates (l)
  - $P(s)P(c|s)P(l|c)$
- light switch (s) $\leftarrow$ closed circuit (c) $\leftarrow$ LED illuminates (l)
  - $P(l)P(c|l)P(s|c)$

Some argue that the causal direction is often simpler to express (functionally less complex), potentially allowing identification of the causal structure from observation alone (Schöllkopf et al., 2012).

However, this heavily relies on a range of assumptions.

The *do* operator

If we want our model to help us:

- make decisions
- fully identify causal relationships, e.g.,

  $X \rightarrow Y \rightarrow Z$ versus $Z \rightarrow Y \rightarrow X$

we need to distinguish between observations and interventions.

Enter the *do* operator.

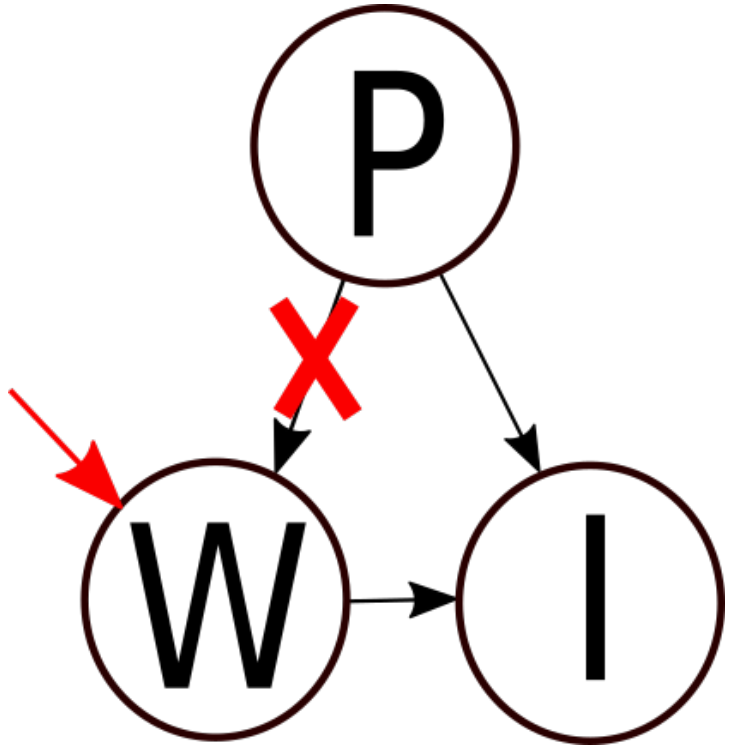
The *do* operator

Intuition: $do(x)$ means "change x, affecting nothing else".

More formally: Mutilate (yes, mutilate) the graph by removing edges from x's parents to x. *Then* condition on x.

If our graph is correct and we know the probability distributions, this **causal graphical model** allows us to make inferences from any combination of observations and interventions.

The *do* operator



Causal graphical models

With this machinery, we can automate causal discovery.

Very difficult in practice due to

- unknown latent variables
- unknown probability distributions
- the large number of possible causal graphs underlying N variables:
  - $N^2 - N$ edges $\rightarrow 2^{N^2-N}$ graphs
  - Event removing cycles, $N = 10 \rightarrow |G| = 4.2E18$

Causal model theories

We can also ask "do people make causal inferences that are consistent with using the machinery of causal graphical models"?

Like Tenenbaum's model of concepts, this is a *rational* model: It predicts the behavior of an agent that behaves in an optimal way, according to a particular definition of optimality, given certain assumptions.

Causal model theories

Causal graphical models can give predictions about causal structure from independence information alone!

Causal identification is usually only possible with interventions or other constraints.

However, testing for statistical dependence is very data intensive whereas people make causal judgments quickly.

Causal model theories

A Bayesian approach can be as *sample-efficient* as people are, but requires assumptions about $P(v_i|pa_i)$.

The standard assumption for *generative* causes is that they are individually sufficient to bring about their effects, but can sometimes fail (independently).

Noisy-OR

This is the "noisy-OR" parameterization:

$$P(v_i|pa_i) = 1 - \prod_{j \in pa_i} (1 - w_j)$$

where w_j is the weight or *causal power* of a cause, assuming it is present or active.

For example, if three causes are present and all have a weight of 0.75, the probability of the effect is $1 - 0.25^3$, or $63/64$.

Noisy-OR

Cheng (1997): People made judgments consistent with MLE estimates of $\mathbf{w}$ under a noisy-OR causal graphical model.

(Though it was Glymour (1998) who framed her results in these terms)

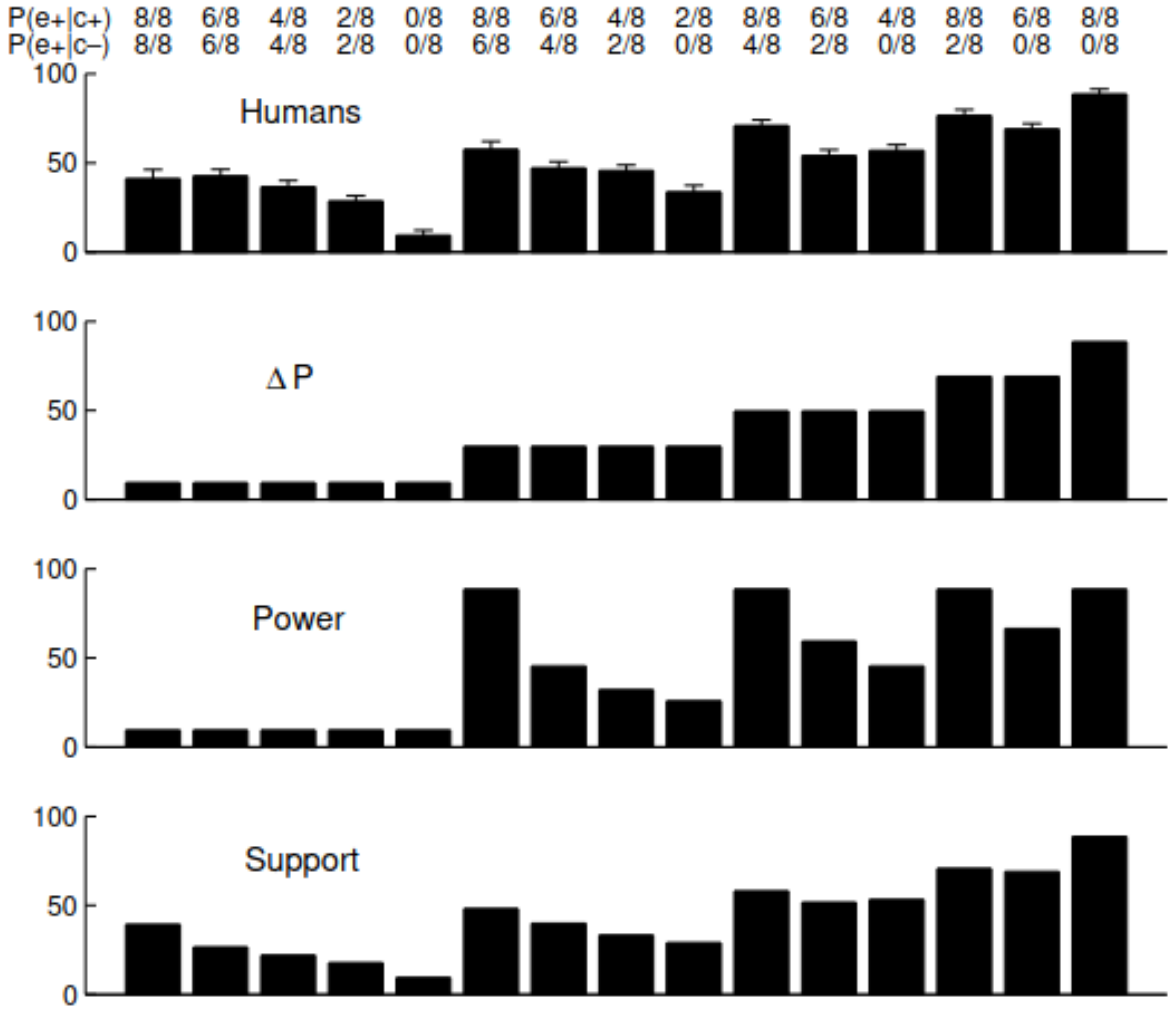
Causal model theories

Since Cheng (1997), there have been many developments, including models and data related to

- The human ability to learn about causal structure as well as strength
- People's expectations (priors) about causal relationships
- Active learning of causal relationships
- Learning more complex causal relationships
- Time and causality

Causal model theories

Example: A model of how people learn structure and strength.



(Clipped from Figure 1 in Griffiths and Tenenbaum, 2005)

Causal model theories

There has also been evidence about the limits of causal model theories, e.g., in explaining

- Order effects that are not predicted by causal graphical models
- Failures to learn structure in the presence of many variables
- Failures to accurately track detailed probabilities

These have inspired new models, most of which build on causal graphical models.

References

- Cheng, P. W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104(2), 367–405. https://doi.org/10.1037/0033-295X.104.2.367
- Glymour, C. (1998, July). Psychological and normative theories of causal power and the probabilities of causes. In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (pp. 166-172).
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. Cognitive psychology, 51(4), 334-384.