

Causal Inference

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 9 of "The NeuroAI Handbook"

Telling "this caused that" from "these just happened together," in the brain and in the machine.

One-sentence thesis. Correlation is what you see; causation is what happens when you intervene, and everything in this chapter is machinery for getting the second when you can only measure the first.

PART 1: Seeing Is Not Doing

Did This Cause That, or Did They Just Co-occur?

A neuron in visual cortex and one in motor cortex fire together. Did the first drive the second? You cannot tell from watching, because watching only reveals that they *co-occur*. The one clean way to know is to **intervene**: force the visual neuron off and see whether the motor neuron changes. That is the interventional definition of cause, "A causes B" means that making A different makes B different, and it is why the **randomized controlled trial** is the gold standard: randomizing who gets the treatment severs every hidden influence on who gets it.

Judea Pearl organized this into a three-rung **ladder of causation**, and the rungs are not interchangeable:

$$\underbrace{P(Y | X)}_{\text{see}} \neq \underbrace{P(Y | do(X))}_{\text{do}}$$

Rung one, **association**, is $P(Y | X)$, passively observing X . Rung two, **intervention**, is $P(Y | do(X))$, actively *setting* X and cutting whatever normally causes it. Rung three, **counterfactual**, asks what would have happened to *this* subject had X been different. No amount of rung-one data answers a rung-two question on its own, you need an intervention or a causal model. The whole discipline lives in that gap between seeing and doing.

The Three Ways a Correlation Is Born

Two variables can move together for reasons that have nothing to do with one causing the other. Every spurious correlation traces to one of three graph shapes. In a **chain** ($X \rightarrow Z \rightarrow Y$), X acts through a mediator. In a **fork** ($X \leftarrow Z \rightarrow Y$), a **confounder** Z drives both, and X and Y dance together while neither touches the other. In a **collider** ($X \rightarrow Z \leftarrow Y$), two independent causes feed a common effect.

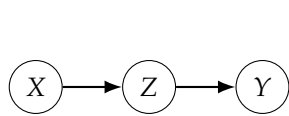
The confounder is the classic trap. Two neurons cor-

relate not because one drives the other but because a presented stimulus drives both, the stimulus is Z . Control for the right variables and the illusion dissolves; miss one and your estimate stays biased no matter how much data you pile up.

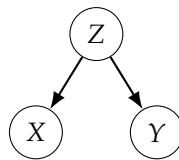
Watch out, colliders bite the other way. Conditioning on a chain or fork *removes* a spurious link, but conditioning on a collider *creates* one. Filter your data by a common effect and you manufacture a correlation between its causes that was never there.

ANALOGY

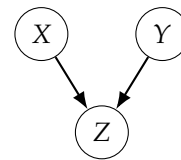
Roosters crow just before sunrise, every day, a near-perfect correlation. Silence every rooster and the sun still rises: the earth's rotation is the confounder driving both. Correlation this strong, and still zero causation.



Chain (mediator)



Fork (confounder)



Collider (common effect)

The three elementary structures. Conditioning on Z blocks the chain and the fork (good, it removes spurious flow) but opens the collider (bad, it invents a correlation between X and Y).

When You Cannot Intervene

Often you are stuck with observational data, no laser, no randomizer. You are not helpless, but every method now rests on assumptions you must defend.

The first tool is **regression with confounding control**: put the confounders into the model alongside the variable of interest and let it isolate the unique relationship. **The catch.** It only works if you measured *every* important confounder. Leave one out and the estimate is still biased, and the regression cannot warn you.

The subtler tool is an **instrumental variable (IV)**, a lever that jostles the treatment but touches the outcome *only* through it. A valid instrument must satisfy three conditions:

- (1) relevance: linked to X
- (2) exclusion: affects Y only via X
- (3) independence: unlinked to confounders

To study dopamine's effect on learning, a genetic variant altering dopamine-receptor density can serve: it shifts dopamine signaling (relevance), acts on learning only through dopamine (exclusion), and is assigned at conception, before any confounder (independence). Because the instrument is as-good-as-random, it recovers a causal effect even with unmeasured confounders. Finally, **causal discovery** algorithms like PC (Spirtes and Glymour) run the logic in reverse, sifting patterns of conditional independence to propose which DAGs could have generated the data.

PART 4: Causality in the Lab and the Machine

The Same Idea in Neurons and Networks

The power of the interventional view is that it unifies a wet lab and a GPU. In neuroscience, **optogenetics** is a physical *do*-operator: light activates or silences a targeted population with millisecond precision, cutting the incoming arrows and revealing $V1 \rightarrow \text{Motor}$ directly, exactly the surgery the graph on the right performs.

In deep learning, **network ablation** is the mirror image, zero out a unit or layer and measure how the

prediction shifts to gauge that component's causal contribution, the AI analogue of a lesion study. This matters beyond curiosity: a model that leans on **spurious correlations** in its training data, snow in the background rather than the animal, shatters under distribution shift. **Causal representation learning** (Schölkopf and colleagues) aims to make models grab the stable, causal features instead, so they keep working when the world changes. Understanding *why* a model decides, not just that it is accurate, is what makes it robust and safe to deploy.



Left: observation leaves the confounding path $X \leftarrow Z \rightarrow Y$ intact, so the X - Y association is contaminated. Right: intervention (optogenetics, or ablation in a net) cuts the arrow into X , isolating the true $X \rightarrow Y$ effect. This is what an RCT buys you.

Structure	Shape	Correlation without causation?	Condition on Z
Chain	$X \rightarrow Z \rightarrow Y$	Real effect, via mediator	Blocks the path
Fork	$X \leftarrow Z \rightarrow Y$	Yes, confounding	Blocks spurious link
Collider	$X \rightarrow Z \leftarrow Y$	No, until you select on Z	Opens a false link

Quick Self-Test

1. State the interventional definition of “A causes B,” and explain why a randomized controlled trial satisfies it.
2. Two neurons fire together during a task. Give a confounding explanation and describe the optogenetic experiment that would settle whether one drives the other.
3. Why does conditioning on a collider create a correlation, whereas conditioning on a confounder removes one?
4. List the three conditions an instrumental variable must satisfy, and say which one lets it beat unmeasured confounding.
5. How is zeroing out a layer in a neural network the same operation as the *do*-operator, and why does it help build robust models?

If You Remember One Thing

$P(Y | X)$ and $P(Y | \text{do}(X))$ are different quantities, and no pile of observational data collapses that difference on its own. Whether you reach for a randomized trial, an instrument, an optogenetic laser, or a network ablation,

causal inference is the art of manufacturing the “do” when the world will only let you “see.”

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552