Your Causal Parrot Might Be Lying To You

and what you can do about it

Prasanna Bhogale
prasanna@romulan.ltd

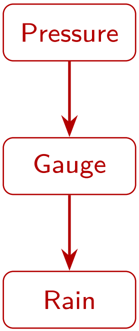
The Fifth Elephant 2025 Winter Edition

ROMULAN

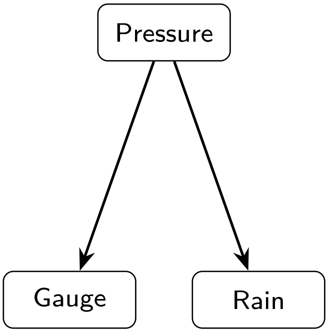
What is in here?

1. AI, BI, why CI?
2. Why can't I just ask Claude?
3. Okay, so how do I do this?
4. Claude gets to play a role after all!

Humans do causal inference effortlessly



Wrong

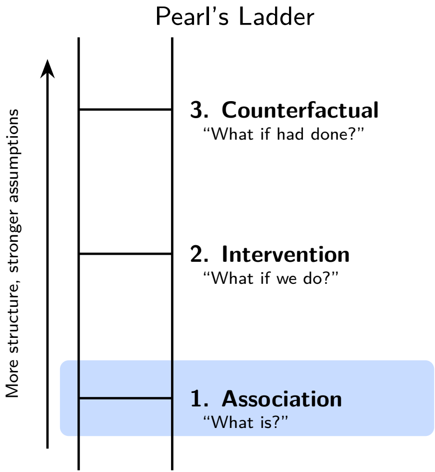


Correct

Your first DAGs!

Human intuition + associational data = good decisions

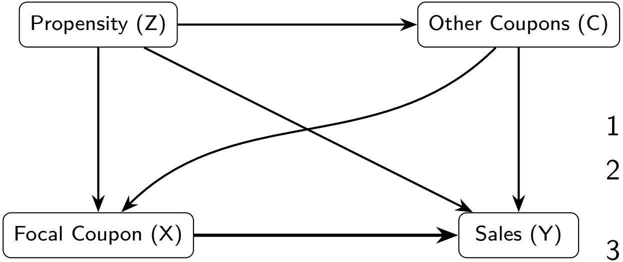
...most of the time



Dashboards, ML, LLMs live on Rung 1

- BI dashboards → **correlations**
- ML models → **patterns**
- **You** bring the causal model (in your head)

Sometimes.. intuition isn't enough



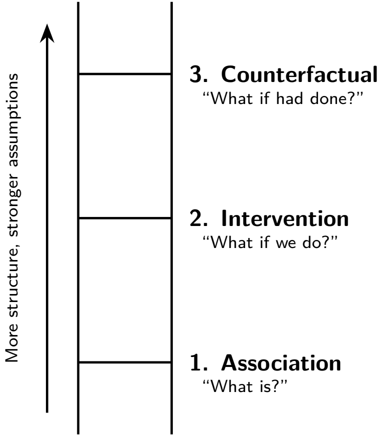
Did coupon *cause* sales,
or just target likely buyers?

1. **New territory** — weak priors
2. **Disagreement** — marketing vs. finance
3. **High stakes** — can't experiment

Q. When will an LLM intuit causality correctly while human experts disagree?

Pearl's Ladder of Causation

Pearl's Ladder

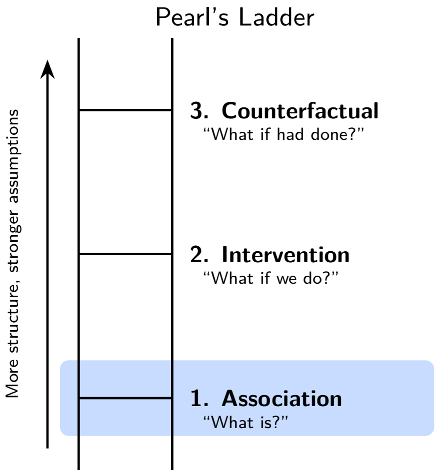


Each rung adds structure:

- **Rung 1:** Add DAG (assumptions)
- **Rung 2:** Interpret edges as causal
- **Rung 3:** Add structural equations

Pearl & Mackenzie, *The Book of Why* (2018)

Rung 1: Association — "What is?"



Observe patterns and correlations

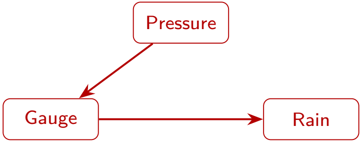
- $P(Y|X)$ — probability given we *observe*

ADD: DAG encoding assumptions

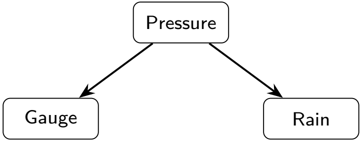
GET: Testable implications

- Different DAGs $\rightarrow$ different conditional independencies
- The DAG is a *falsifiable* hypothesis.

Rung 1: The DAG is a *falsifiable* hypothesis.



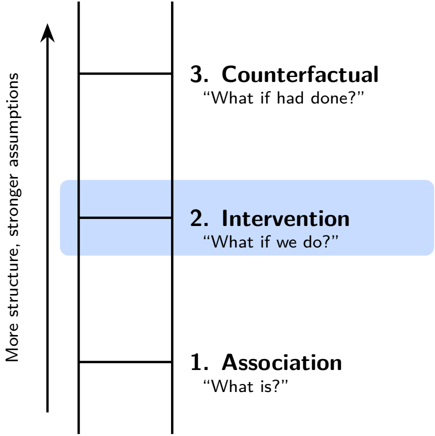

```
impliedConditionalIndependencies('
 dag {Pressure -> Gauge;
 Pressure -> Rain}')
# Gaug _||_ Rain | Prss
```




```
impliedConditionalIndependencies('
 dag {Pressure -> Gauge;
 Gauge -> Rain}')
# Prss _||_ Rain | Gaug
```

Rung 2: Intervention — "What if we do?"

Pearl's Ladder



Imagine (or perform) interventions

- $P(Y|\text{do}(X))$ — probability if we *force* X
- $\text{do}(X) \neq$ observe X **(key insight!)**

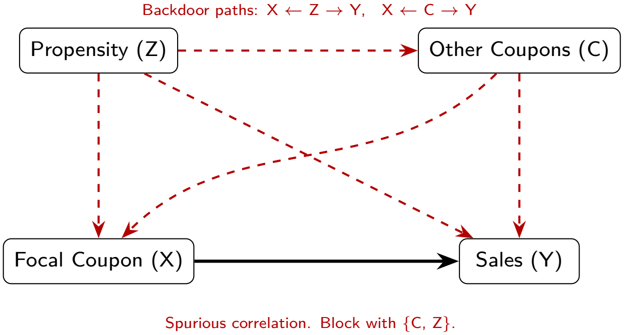
ADD: Interpret edges as *causal*

GET: Causal effect estimation

- "Graph surgery" — intervening severs incoming arrows
- Estimate from observational data (if identifiable)

The Backdoor Criterion — the key to Rung 2

A set of variables Z satisfies the back-door criterion for estimating the causal effect of X on Y if no variable in Z is a descendant of X and Z blocks every path from X to Y that starts with an arrow into X.



Problem:

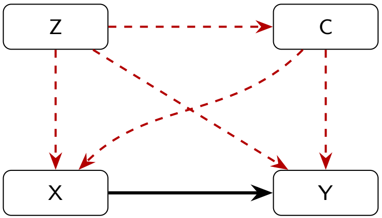
- Want $P(Y|\text{do}(X))$, only observe $P(Y|X)$
- *Not the same* with confounding!

Non-causal path $\rightarrow$ spurious correlation

The Backdoor Criterion — the key to Rung 2

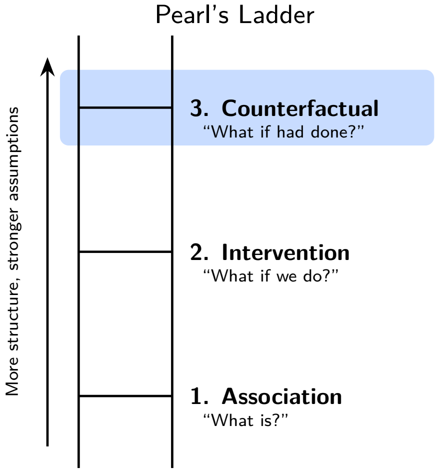
```
dag <- dagitty('dag {
  Z -> X; Z -> Y;
  C -> X; C -> Y;
  X -> Y}')

adjustmentSets(dag,
 exposure = "X", outcome = "Y")

# Result: { C, Z }

dosearch("P(X, Y, Z, C)", "P(Y | do(X))", dag)
```

$$\sum_{C,Z} p(C,Z)p(Y|X,C,Z)$$



Answer: Adjust for $\{C, Z\}$

Rung 3: Counterfactual — "What if we had done?"



Specific individuals, alternative histories

"What if *this customer* hadn't got the discount?"

ADD: Structural equations (functional forms)

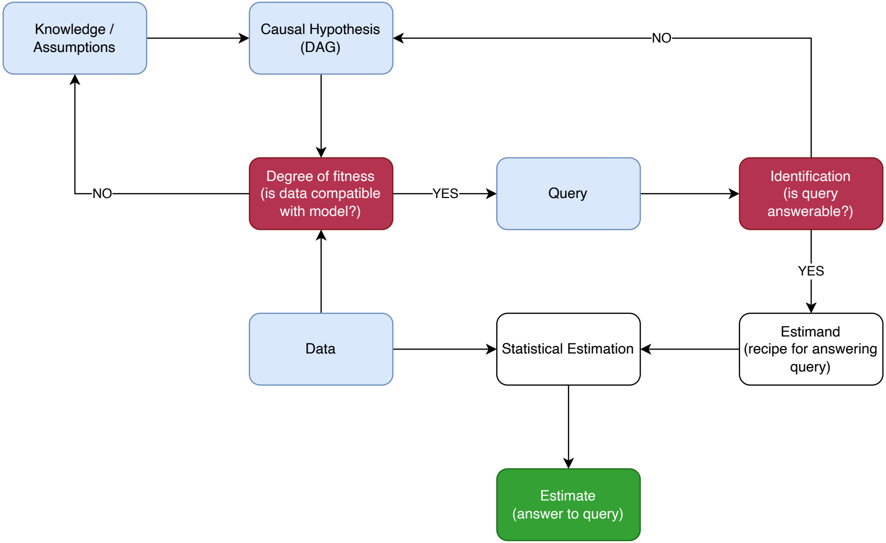
GET:

- Individual-level attribution
- Regret analysis, fairness

Hardest rung — strongest assumptions.

Focus today: Rungs 1 & 2

The "artisanal" causal Inference framework



LLM failure modes = "New Intern" failure modes

Okay Claude..

LLM failure modes = "New Intern" failure modes

Okay Claude..

LLM tells you about $P(Y|X)$, cannot tell you about $P(Y|\mathrm{do}(X))$

CLadder Benchmark — LLMs get worse as we climb the ladder

Model	Overall	R1	R2	R3
Random	49%	50%	48%	49%
GPT-4	62%	63%	63%	60%
GPT-4 + CoT	**70%**	83%	67%	**62%**

Jin et al., NeurIPS 2023

Anti-commonsensical scenarios:

- Causal relationships $\neq$ "internet wisdom"
- Performance drops further
- $\Rightarrow$ Pattern-matching, not reasoning

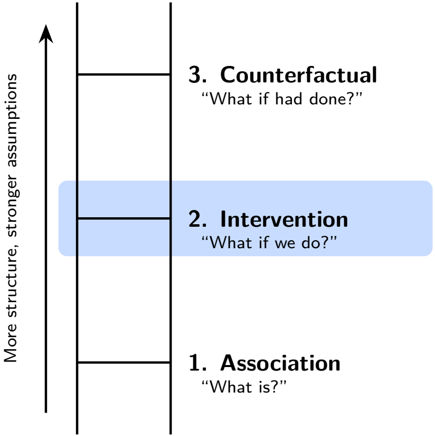
Okay, lets do this then..

Estimated causal effect (Rung 2)

```
lm(Y ~ X)          # Unadjusted
lm(Y ~ X + C + Z)  # Adjusted
```

Coupon Type	Naive	Adjusted
Drugstore	29.94	96.57
Ready-to-eat	25.95	21.30

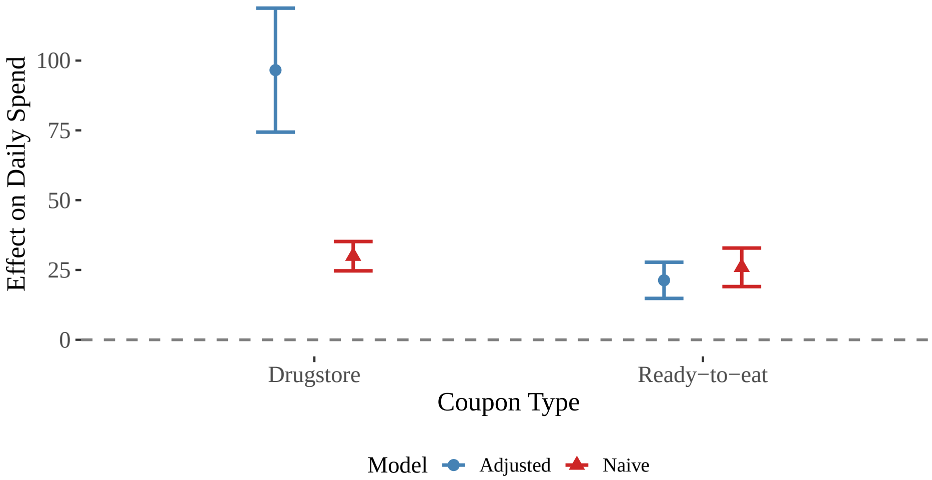
Pearl's Ladder



This is what an LLM gets wrong.

Naive vs Adjusted Effect Estimates

Adjusting for confounders {Z, C} changes effect estimates

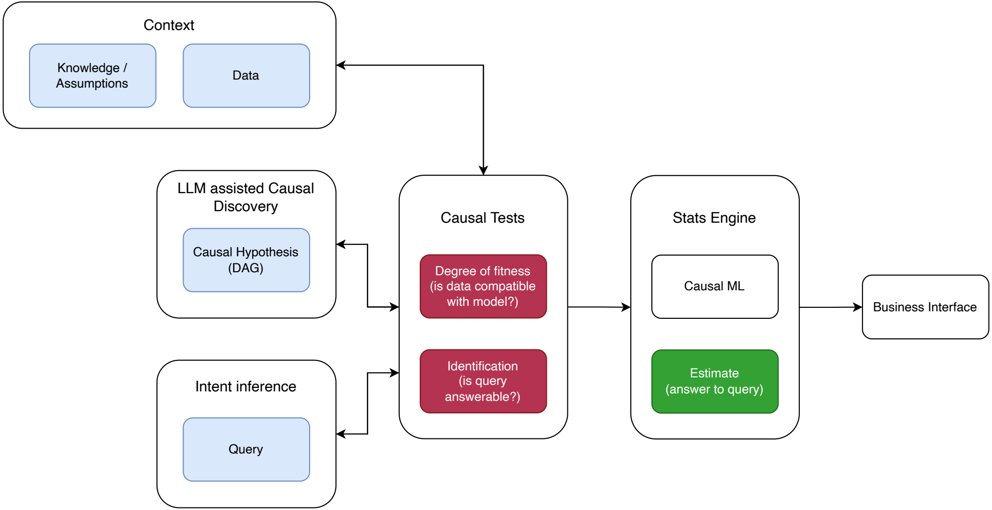


What we just did

1. Knowledge $\rightarrow$ Assumptions $\rightarrow$ DAG
2. DAG $\rightarrow$ Testable implications
3. Query $\rightarrow$ Answerable? $\rightarrow$ Adjust!
4. Data $\rightarrow$ Estimate

Explicit. Auditable. Defensible.

"Not-so-artisanal" causal inference



Limitations

- Some questions are **genuinely unanswerable**
- Sometimes we **can't know the right DAG**
- Sometimes we **lack measurements**

Explicit assumptions > Hidden assumptions

Takeaways

1. **Pearl's Ladder:** Seeing $\rightarrow$ Doing $\rightarrow$ Imagining
2. **DAGs** $=$ explicit, auditable, defensible, causal assumptions
3. **LLMs** $=$ Rung 1 only (causal parrots)
4. **Solution:** LLMs for hypotheses, tools for estimation

We'd love to chat about your hard causal (or other) problems!

`prasanna@romulan.ltd`

More at: `https://theclarkeorbit.github.io/`

Questions?

Backup slides

LLMs: Powerful interns, but still parrots

Good at:

- Summarizing patterns
- Identifying variables
- Drafting causal stories
- Brainstorming DAGs

Fail at causal reasoning:

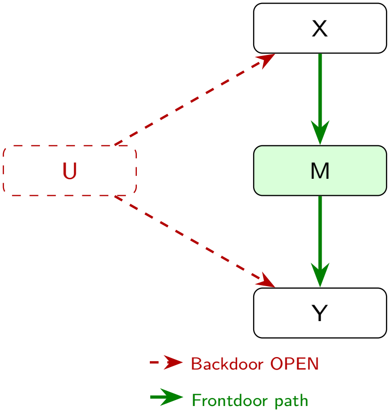
- Pattern-match on internet language
- Lack *your* business DAG
- No do-calculus machinery
- Can't know when wrong

"LLMs may just be 'causal parrots'"

— Zečević et al., TMLR 2023

Great for drafting, dangerous for deciding.

Front Door Criterion



When backdoor criterion fails:

- Backdoor $X \leftarrow U \rightarrow Y$ is **open**
- Can't condition on U (unmeasured!)
- But: mediator M is observed

Frontdoor works if:

1. M intercepts all directed $X \rightarrow Y$ paths
2. No unblocked backdoor $X \rightarrow M$
3. All $M \rightarrow Y$ backdoors blocked by X

Classic: Smoking → Tar → Cancer (genetics unmeasured)

Causal Identification: The General Problem

Given:

- Causal DAG (assumptions)
- Observational distribution $P(V)$
- Query: $P(Y|\text{do}(X))$

Question:
Can we express the query using only observational data?

If yes $\to$ **identifiable**
If no $\to$ need experiment or more assumptions

Identification strategies:

1. Backdoor criterion
2. Frontdoor criterion
3. Instrumental variables
4. **do-calculus** (complete)

Tools:

- `dagitty::adjustmentSets()`
- `dosearch::dosearch()`
- `causaleffect` (R package)

Algorithms can determine identifiability automatically!

Causal ML: Heterogeneous Treatment Effects

Beyond ATE:
Average effect hides variation

CATE: Conditional Average Treatment Effect

$$\tau(x) = E[Y(1) - Y(0)|X = x]$$

Question:
Who benefits most from treatment?

Personalization, targeting, policy optimization

Methods:

- **Causal Forests** (Athey & Wager)
- Double/Debiased ML (Chernozhukov)
- Meta-learners (S, T, X-learner)
- Bayesian approaches

R packages:

- `grf` (causal forests)
- `DoubleML`
- `causalweight`

Still need valid identification (DAG) first!