

LECTURE 12

Causal, mechanistic, and counterfactual reasoning

Observational, interventional and counterfactual queries.

Reasoning agents, feedback, and scientific evidence

Prediction leaves an intervention unresolved

A measured variable X predicts an outcome Y . That is useful for forecasting, but it does not determine what happens if we actively change X .

Both could respond to another variable Z . Alternatively, changing X could itself change Y . The same association can support different intervention decisions.

A practical intervention makes the ambiguity consequential. We need a representation of how the observations could have been generated.

Constructing competing causal explanations

Construct one model in which Z affects both X and Y , and another in which X affects Y . Ask which observations could distinguish them.

An intervention replaces the rule generating X while leaving the other structural rules in place. This operation makes the causal question precise.

Adjustment, a new measurement or a randomized intervention can then be considered under explicit assumptions. The graph itself remains a scientific claim to justify.

Choosing evidence for a causal question

Candidate	Reason to choose it	Cost or limitation
Prediction from association	Forecasting under a stable observed regime	Does not identify an intervention effect
Adjustment using measured causes	A defensible graph and sufficient data exist	Requires identification, consistency and positivity
Controlled intervention	Manipulation and measurement are feasible	May be costly, constrained or context-specific

Choose evidence for the causal query and its assumptions. More predictive accuracy cannot resolve every ambiguity between causal models.

Pedagogical reconstruction

[Pearl \(2009\)](#)

A structural causal model specifies mechanisms

$$X_i = f_i(\text{Pa}_i, U_i)$$

do($X = x$) replaces the equation for X

Each variable is generated by a function of its parents and an exogenous disturbance. A directed graph summarises the assumed dependencies.

An intervention replaces the structural equation for the intervened variable while retaining the other mechanisms. This modularity is an assumption, not a consequence of observing correlation.

Measurement error, hidden confounding, and context changes can undermine the model's interpretation.

Adjustment blocks the relevant backdoor paths

$$p(y \mid \text{do}(x)) = \sum_z p(y \mid x, z) p(z)$$

When a measured variable Z confounds X and Y and satisfies the adjustment conditions, average the conditional outcome distribution over the population distribution of Z .

The formula uses $p(z)$, not the treatment-specific composition $p(z|x)$. It requires adequate support and valid causal assumptions.

Adjusting for every available variable can introduce bias, particularly when conditioning on a collider or a mediator inappropriate for the target effect.

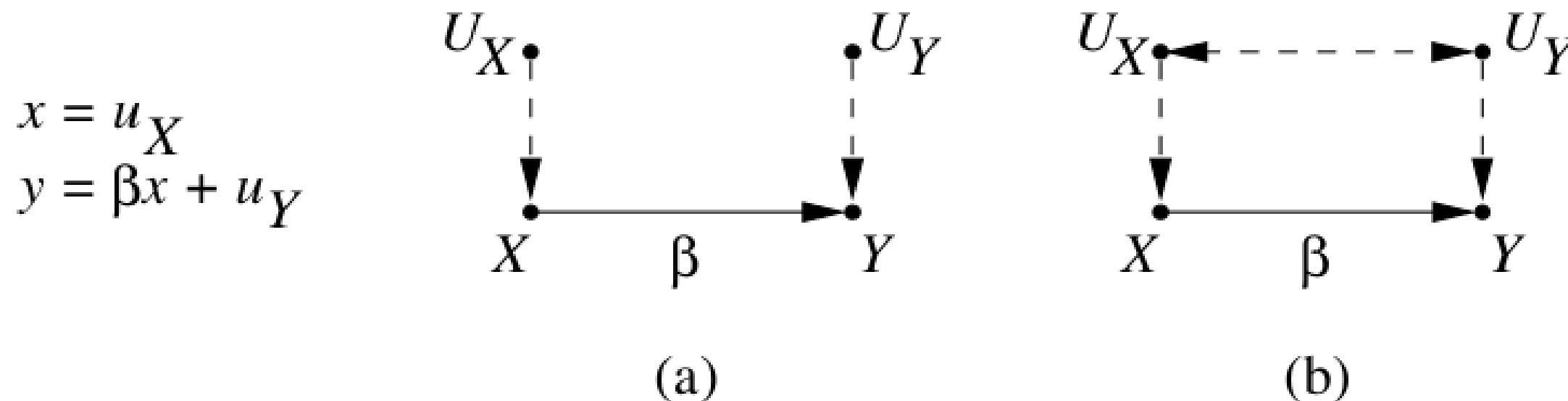
Structural causal models: the framework

Pearl's structural causal framework represents variables as functions of parents and exogenous disturbances, and interventions as modifications to those mechanisms.

Graphical assumptions support identification of causal quantities from suitable data. Counterfactual queries add unit-specific inference about exogenous conditions.

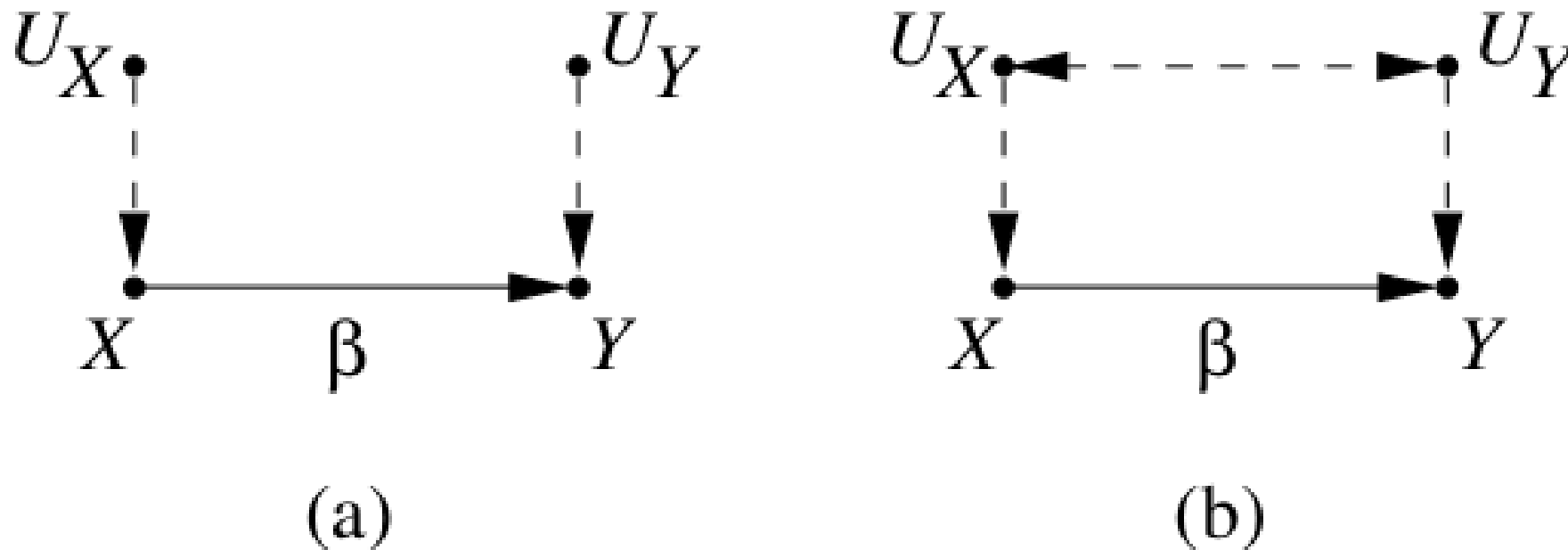
The framework distinguishes what is assumed, what is identified, and what must be estimated, preventing an association or a predictive fit from automatically becoming a causal claim.

Structural causal models: illustrative assumptions



The simple structural model uses independent disturbances; the alternative allows correlated disturbances. The dashed connection changes the relation between observed covariance and the causal coefficient. These are illustrative models, not measured effects in a named dataset.

Causal assumptions about exogenous disturbances



The exogenous-variable dependence differs between the two graphs. Interpreting covariance as the direct causal coefficient requires assumptions about those disturbances.

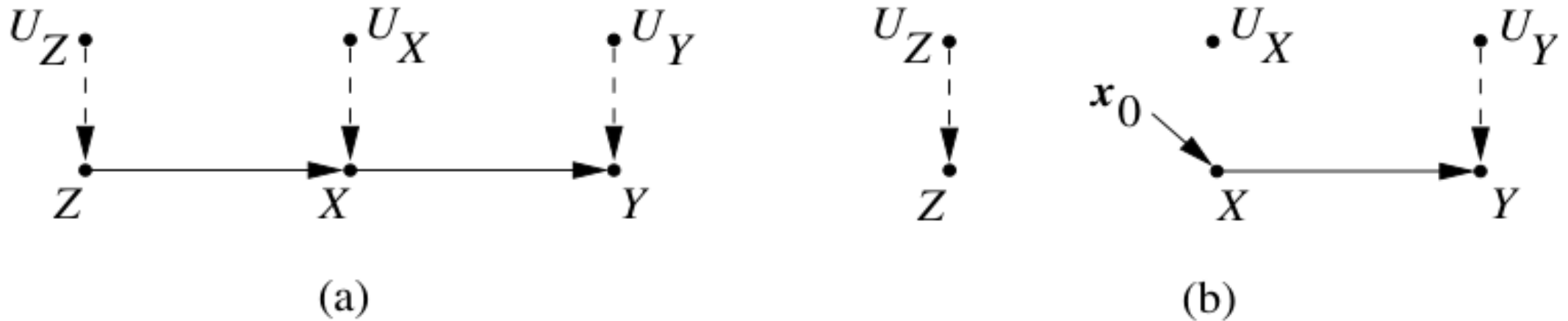
Evaluation scope

The intervention diagram removes the incoming causal mechanism for the variable that is set externally while preserving the other equations.

The unchanged equations encode a modularity assumption. The resulting interventional distribution can differ from conditioning on the same observed value.

The paper develops identification and adjustment rules; their use requires checking the assumptions and data support relevant to the target query.

Structural causal models: interpretation and limits



The graph is an assumption about the data-generating process. Hidden confounding, measurement error, selection, and incorrect invariance assumptions can invalidate a causal interpretation even when the algebra is correct.

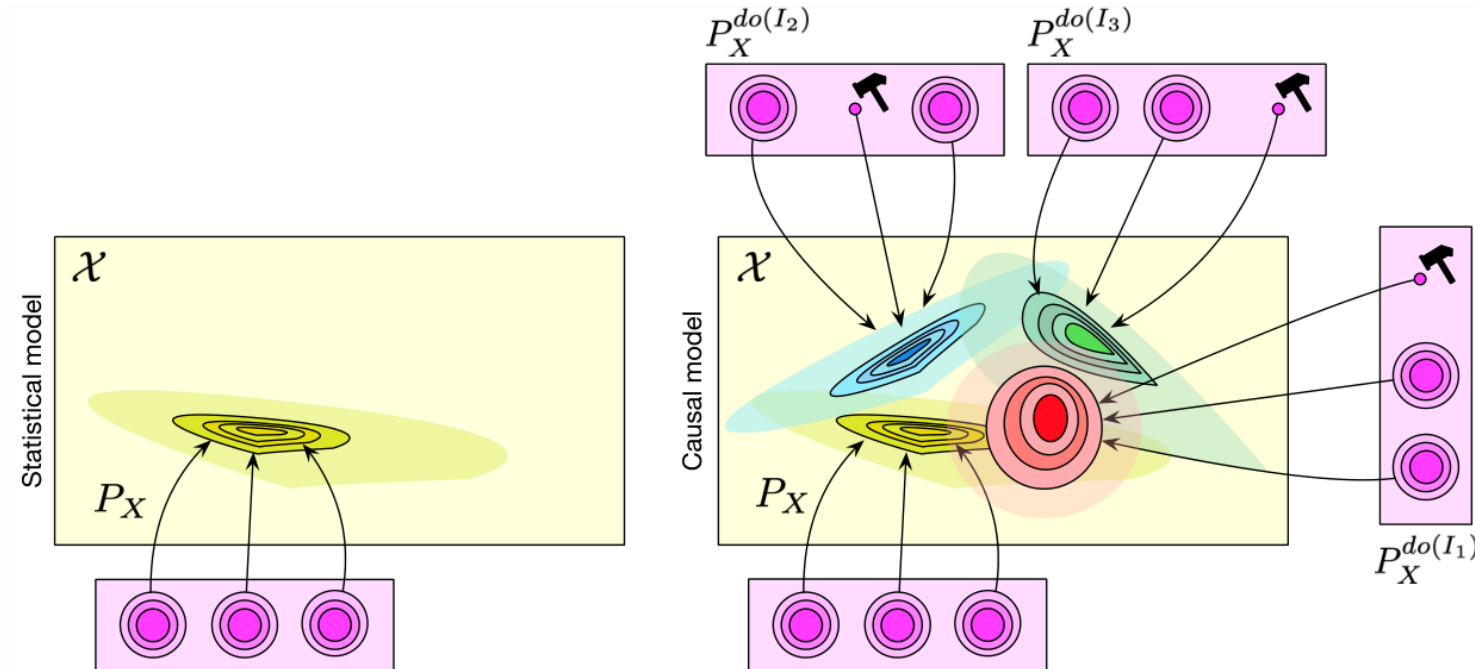
Causal representation learning: research framework

The paper examines how learned representations might capture causal variables and mechanisms rather than only statistical regularities in observations.

It relates interventions, distribution changes, independent mechanisms, and representation learning.

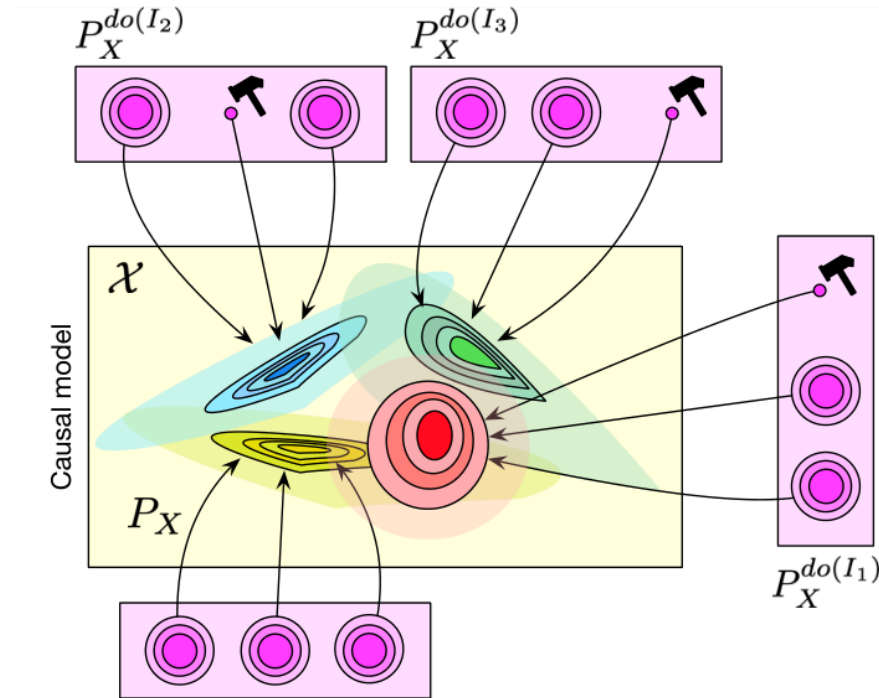
The goal is to support transfer and reasoning about changes to a system, while recognising that observational data alone often leave representations or causal structures unidentified.

Observational and interventional distributions



A statistical distribution describes observations, while a causal model also specifies interventional distributions. Additional assumptions or data connect observational density to predictions under intervention.

Causal models and changes to mechanisms



The interventions change a specified mechanism while preserving relations implied by the causal model. This defines a narrower transfer claim than a generic assertion of robustness.

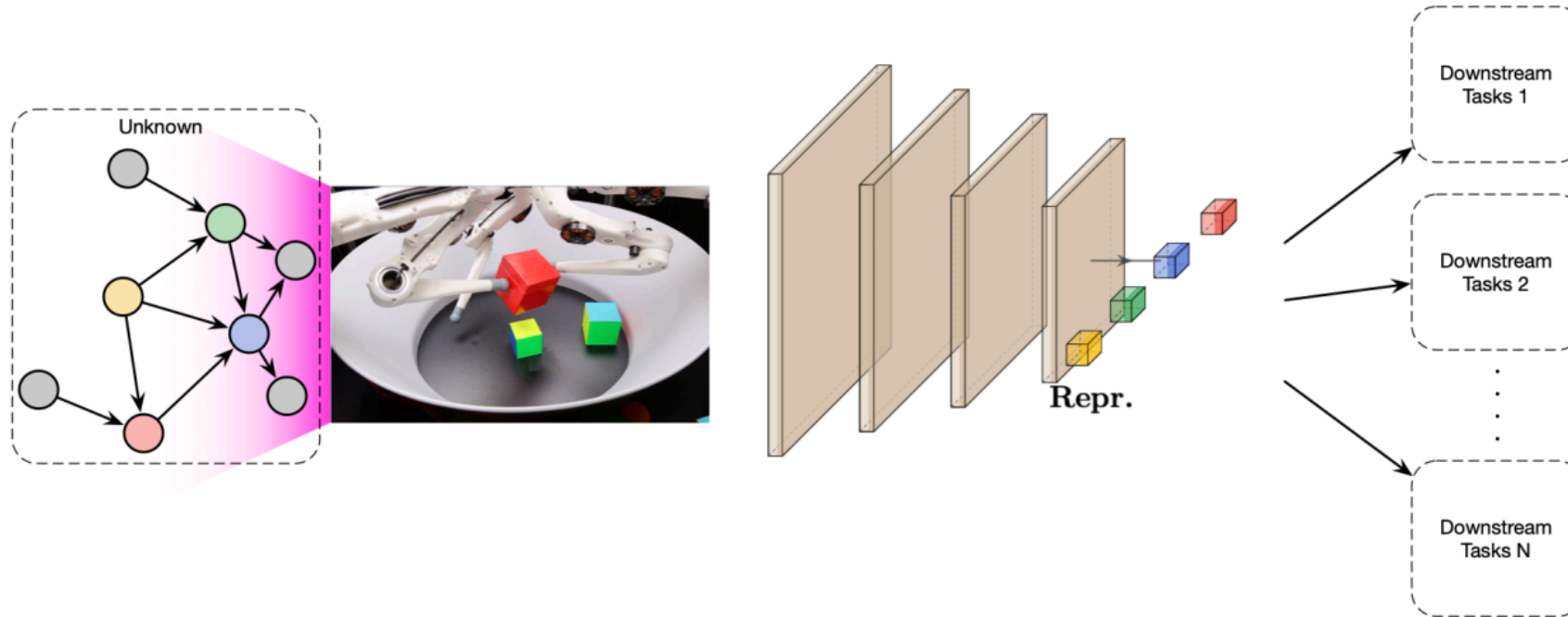
Evaluation scope

The representation-learning diagram links unknown causal structure, high-dimensional observations, a learned representation, and downstream tasks.

It is a conceptual research framework rather than a controlled benchmark demonstrating one universally identified causal representation.

Read each proposed principle in relation to the data and intervention assumptions it requires. Generalisation benefits must be evaluated under the distribution shifts relevant to the application.

Causal representation learning: interpretation and limits



Identifiability requires assumptions or additional information. A disentangled or compressed representation is not automatically causal, and good predictive transfer does not uniquely identify the underlying mechanisms.

Counterfactual reasoning has three stages

Abduction uses observed evidence to infer exogenous conditions for the unit. Action changes the relevant structural equation. Prediction propagates those inferred conditions through the modified model.

A counterfactual estimate can require stronger assumptions than an average interventional effect. Observational fit alone may leave it unidentified.

State which model features are learned from data and which are supplied by scientific knowledge.

A causal effect leaves a mechanism question open

Suppose changing X changes Y. That rejects the claim that X has no effect, but does not tell us which intervening process carries the effect.

Construct two candidates: X acts directly on Y, or X changes mediator M and M changes Y. Both can predict the same response to changing X.

The new direction is to intervene inside the proposed process. More detailed stories help only when they make distinct, testable commitments.

Language models can propose causal structure

An LLM can retrieve relevant knowledge, propose a graph, identify possible confounders, and translate a query into formal notation. Each step may be useful and fallible.

Evaluate factual knowledge separately from valid causal inference. A benchmark with known causal descriptions may not test discovery of unknown mechanisms from data.

Validate proposed edges, adjustment sets, and interventions using domain evidence and explicit assumptions before using the resulting estimate.

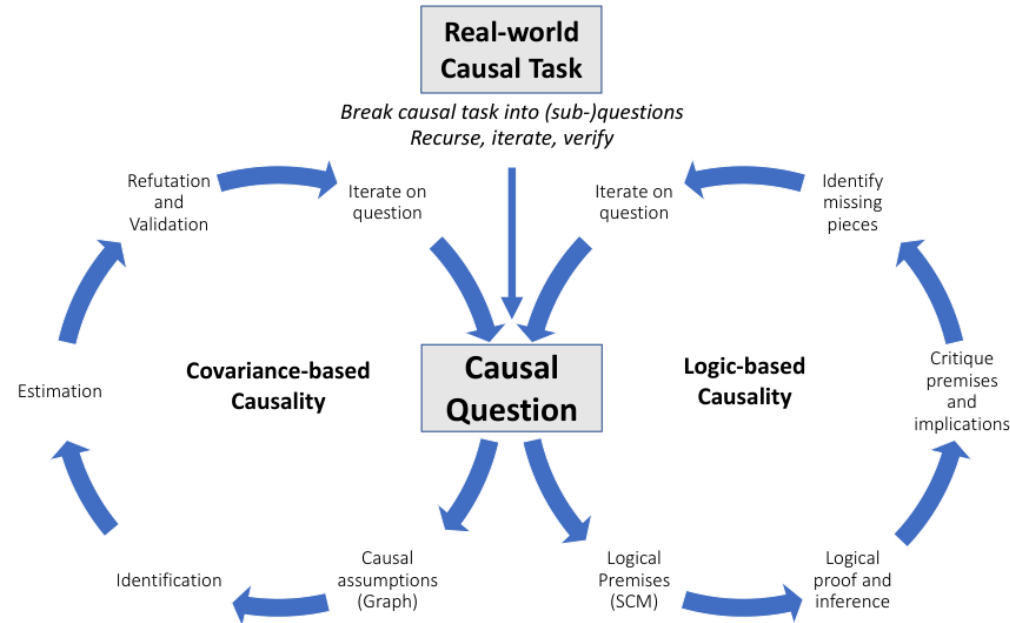
Language models in causal reasoning: study scope

The study explores language models as sources of causal knowledge and as assistants for constructing and analysing causal arguments.

Tasks include identifying cause–effect directions and reasoning with descriptions of variables or causal situations.

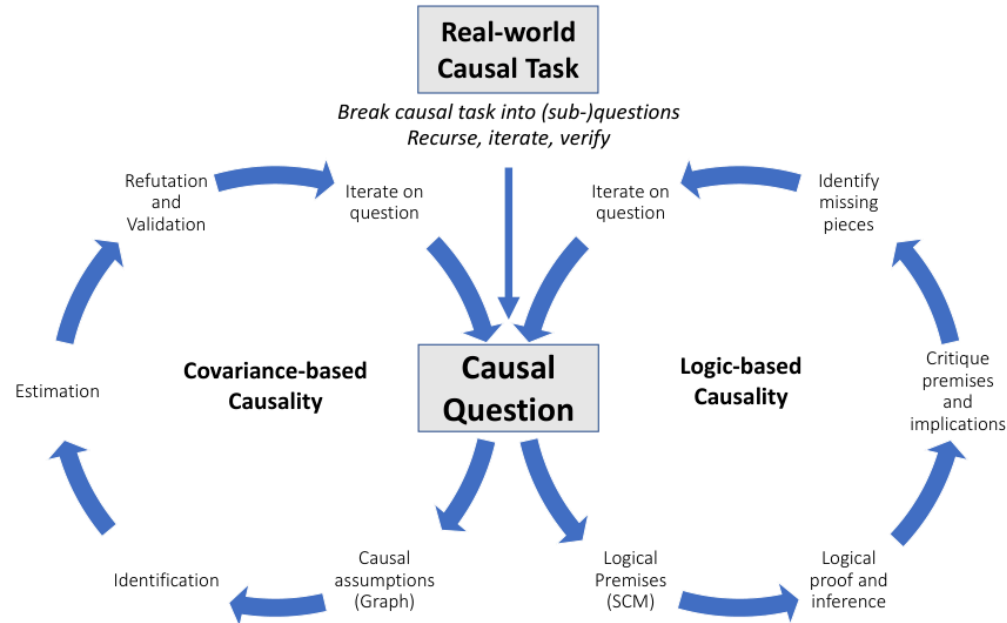
Performance can depend on prior factual knowledge and language cues, so the evaluation must distinguish knowledge retrieval from discovery of unknown causal structure using data.

Causal reasoning: data, knowledge and inference



The diagram relates covariance-based and logic-based causal inference. Causal work combines data analysis with knowledge, assumptions and formal reasoning. Assistance at several stages does not independently validate the whole chain.

Causal reasoning: assumptions before inference



The transition from a causal question to formal inference requires assumptions. Facts recalled by a model and claims justified by data or experiments have different evidential roles.

Evidence: information access and accuracy

Model	Acc.	Wt. Acc.
Slope (Marx & Vreeken, 2017)	0.75	0.83
bQCD (Tagasovska et al., 2020)	0.68	0.75
PNL-MLP (Zhang & Hyvarinen, 2012)	0.75	0.73
Mosaic (Wu & Fukumizu, 2020)	0.83	0.82
ada	0.50	0.50
text-ada-001	0.49	0.50
babbage	0.51	0.50
text-babbage-001	0.50	0.50
curie	0.51	0.52
text-curie-001	0.50	0.50
davinci	0.48	0.47
text-davinci-001	0.50	0.50
text-davinci-002	0.79	0.79
text-davinci-003	0.82	0.83
LMPrior (Choi et al., 2022)	0.83	-
gpt-3.5-turbo	0.81	0.83
gpt-3.5-turbo (causal agent)	0.86	0.87
gpt-3.5-turbo (single prompt)	0.89	0.92
gpt-4 (single prompt)	0.96	0.97

For Tübingen cause-effect pairs, GPT-4 has accuracy 0.96 and weighted accuracy 0.97; Mosaic has 0.83 and 0.82. The LLM is prompted with variable names, drawing on prior knowledge, while the statistical methods analyse observations. Identifying a familiar causal direction is distinct from estimating an effect or identifying a new mechanism.

Causal reasoning with language models: interpretation and limits

Model	Acc.	Wt. Acc.
Slope (Marx & Vreeken, 2017)	0.75	0.83
bQCD (Tagasovska et al., 2020)	0.68	0.75
PNL-MLP (Zhang & Hyvarinen, 2012)	0.75	0.73
Mosaic (Wu & Fukumizu, 2020)	0.83	0.82
ada	0.50	0.50
text-ada-001	0.49	0.50
babbage	0.51	0.50
text-babbage-001	0.50	0.50
curie	0.51	0.52
text-curie-001	0.50	0.50
davinci	0.48	0.47
text-davinci-001	0.50	0.50
text-davinci-002	0.79	0.79
text-davinci-003	0.82	0.83
LMPrior (Choi et al., 2022)	0.83	-
gpt-3.5-turbo	0.81	0.83
gpt-3.5-turbo (causal agent)	0.86	0.87
gpt-3.5-turbo (single prompt)	0.89	0.92
gpt-4 (single prompt)	0.96	0.97

Language descriptions and known relationships can supply decisive clues. Prompt sensitivity, factual errors, and task leakage complicate broad conclusions about causal discovery competence.

Construct rivals with the same total effect

Candidate	Equations in the toy system	After setting $X = 1$
D: direct route	$M = X; Y = X.$	$M = 1$ and $Y = 1.$ M is an accompanying marker.
I: indirect route	$M = X; Y = M.$	$M = 1$ and $Y = 1.$ M carries the effect.
Why both survive	At baseline X is a fair bit. All observed values agree.	Even intervening on X alone does not select the mechanism.

Predicting a total effect is weaker than identifying the internal route that produces it.

Pedagogical reconstruction

[Pearl \(2009\)](#)

[Machamer et al. \(2000\)](#)

A joint intervention separates the routes

Evidence	Direct route D	Indirect route I
Set $X = 1$ and $M = 0$	$Y = X = 1$	$Y = M = 0$
Interpret the contrast	Blocking M leaves the $X \rightarrow Y$ route intact.	Blocking M interrupts the proposed route.
Validity requirement	Intervening on M must not change Y by another route.	Measurement and timing must match the proposed mechanism.

A mechanism claim needs evidence about the route; a transfer claim needs evidence about the new setting.

Pedagogical reconstruction

[Pearl \(2009\)](#)

[Machamer et al. \(2000\)](#)

A shared cause creates predictive association

Consider $X=C+U_x$ and $Y=C+U_y$. Let C , U_x and U_y be mutually independent, zero-mean variables with positive variance.

X predicts Y because both contain C . Under an intervention replacing the equation for X , the equation for Y is unchanged. This model gives zero causal effect of X on Y .

Two questions for the same variables

Question	Operation	Result in this model
Does X predict Y?	Observe X	Positive association
Does changing X change Y?	Replace X equation	No effect on Y
Does changing C change Y?	Replace C equation	Y changes with C

Synthetic structural model. Observational association alone does not select this model over every rival.

A counterfactual preserves the inferred background

Operation	Calculation or state	Consequence
Structural model	$Y=3X+U$	U stays fixed for this unit
Observed unit	$X=2, Y=7$	Infer $U=7-3\times 2=1$
Intervention	Set $X=1$	Replace the X mechanism
Counterfactual	$Y=3\times 1+1=4$	Same inferred U

Synthetic structural model with deterministic equations and an observed unit. A population intervention generally averages over backgrounds instead.

Worked adjustment: a confounded comparison

Context Z	Population weight	$P(Y=1 \mid X=1, Z)$	$P(Y=1 \mid X=0, Z)$
0	0.5	0.8	0.6
1	0.5	0.4	0.2

The synthetic probabilities assume that Z is a sufficient adjustment set and that both treatment values occur in both contexts. Weighting both treatments by the same target population distribution yields the adjusted comparison.

Calculate the adjusted outcome rates

$$E[Y \mid \text{do}(X = 1)] - E[Y \mid \text{do}(X = 0)] = 0.6 - 0.4 = 0.2$$

Average each treatment's conditional rates over the same population distribution of Z .
For $X=1$, the adjusted rate is $0.5 \times 0.8 + 0.5 \times 0.4 = 0.6$.

For $X=0$, the adjusted rate is $0.5 \times 0.6 + 0.5 \times 0.2 = 0.4$. The average causal risk difference is therefore 0.2 under the assumptions.

The calculation does not independently establish that Z captures all confounding.

Show how an observational reversal can arise

Suppose 80% of treated cases are in the harder context $Z=1$, while only 20% of untreated cases are there.

The observed treated rate is $0.2 \times 0.8 + 0.8 \times 0.4 = 0.48$. The observed untreated rate is $0.8 \times 0.6 + 0.2 \times 0.2 = 0.52$.

The unadjusted comparison suggests a negative difference even though the assumed causal effect is positive in each context. Different context composition drives the reversal.

A causal audit for a scientific agent

A causal claim is defined by its estimand, graph, intervention, adjustment strategy and assumptions. Data, literature and modelling choices contribute different parts of its justification.

Sensitivity to alternative plausible graphs, negative controls and experimental comparisons helps characterize uncertainty in the causal conclusion.

Successful code execution and a valid causal conclusion are separate outcomes. A perfectly reproduced calculation can still estimate the wrong quantity.

Counterfactuals retain the inferred exogenous state

$$X := U_x, \quad Y := 2X + U_y, \quad (X, Y) = (1, 3) \implies U_y = 1$$

In this illustrative deterministic structural model, abduction infers $U_y=1$ from the observed unit. The intervention $\text{do}(X=0)$ replaces only the X equation.

Prediction then gives $Y_{(x=0)}=1$ for that same unit. A population intervention would instead average over the distribution of U_y .

The numerical counterfactual depends on the structural equation and disturbance assumptions; those are not established by observing one pair.

Identification precedes an estimator

The adjusted risk difference in the worked example is identified under the assumed confounding structure, consistency, and positivity. Increasing sample size cannot repair a wrong graph or absent treatment support.

First specify the causal query and justify the adjustment set. Then choose an estimator and assess uncertainty.

An LLM can assist each stage, but a fluent derivation cannot supply an unobserved confounder or establish an unsupported invariance assumption.

Observational agreement does not identify a cause

Model	Structural equations	Observed distribution
M1: X causes Y	U is a fair bit; $X = U$; $Y = X$	$P(0,0) = P(1,1) = 1/2$
M2: shared cause only	U is a fair bit; $X = U$; $Y = U$	$P(0,0) = P(1,1) = 1/2$
Any amount of these observations	X and Y always agree	Both models fit exactly

A lecturer-created counterexample. Predicting Y from X perfectly does not distinguish these causal explanations.

An intervention separates the two models

Operation	M1: $Y = X$	M2: $Y = U$
Observe $X = 1$	$E[Y \mid X = 1] = 1$	$E[Y \mid X = 1] = 1$
Set X to 1 externally	Replace $X = U$ by $X = 1$	Replace $X = U$ by $X = 1$
After $\text{do}(X = 1)$	$E[Y] = 1$	$E[Y] = 1/2$

The difference comes from changing one structural equation while holding the others fixed. The validity of that intervention is an assumption.

Recent methods in context

Executable Counterfactuals (ICLR 2026). Uses executable problems to distinguish abduction, intervention and prediction in counterfactual reasoning. Success with a supplied causal program does not identify the causal model of a real-world dataset.

CHIVE (Aug 2026 preprint). Investigates observed model behaviors with counterfactual prompt edits and tests explanations through predictive usefulness. The reported lack of uplift from tested interpretability tools is conditional on this evaluation and tool set.

Model Discovery Agent (preprint; revised 2026). Combines Bayesian model/parameter inference, LLM proposals of model structures and value-of-information experiment choice. Version 4 reports fixing prompt leakage and rerunning benchmarks; comparisons must use the corrected version.

References (1/2)

Judea Pearl (2009). Causal inference in statistics: An overview

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer et al. (2021). Towards Causal Representation Learning

Emre Kıcıman, Robert Ness, Amit Sharma et al. (2023). Causal Reasoning and Large Language Models: Opening a New Frontier for Causality

Yusuf Roohani, Andrew Lee, Qian Huang et al. (2024). BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments

Peter Machamer, Lindley Darden, Carl F. Craver (2000). Thinking about Mechanisms

References (2/2)

Samuel Schmidgall, Xiaokai Zhu, Marian Shaw et al. (2026). Accelerating Scientific Research with Gemini in the Real-World

Aniket Vashishtha, Qirun Dai, Hongyuan Mei et al. (2025). Executable Counterfactuals: Improving LLMs' Causal Reasoning Through Code

Adam Karvonen, Euan Ong, Subhash Kantamneni et al. (2026). Would this change your answer? Evaluating Explanations of LLM Behavior In The Wild with Counterfactual Experiments

Kevin Murphy (2026). Model Discovery Agent: LLM-assisted Bayesian experiment design for data-efficient discovery of mechanistic world models