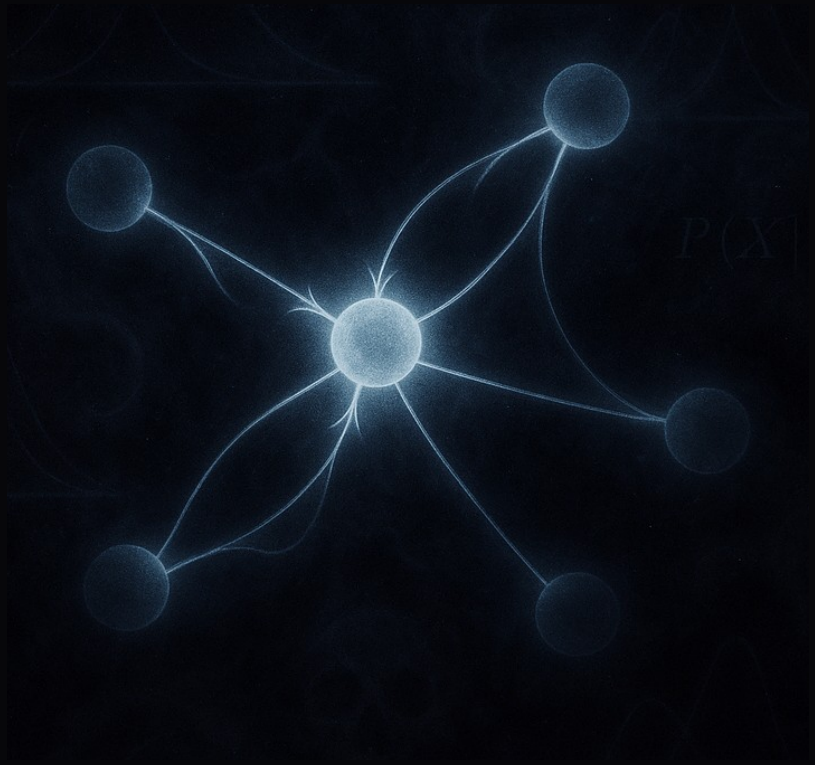


A Bayesian Reduction of Causation

Aydin Mohseni

Department of Philosophy

Institute for Complex Social Dynamics

Carnegie Mellon University

In collaboration with...



Daniel Herrmann

Department of Philosophy

University of North Carolina at Chapel Hill



Ben Levinstein

Alignment Researcher

Anthropic



Gerard Rothfus

Department of Philosophy

Texas Christian University



Bruce Rushing

Research Scientist

University of Virginia

An expression of influence and gratitude to our teachers...

Causal Modeling, Learning, and Reasoning



Peter Spirtes

Marianna Brown Dietrich Professor and Head of Philosophy



Clark Glymour

Senior Research Scientist at IHMC

Alumni University Professor



Richard Scheines

The Bess Family Dean of the Marianna Brown Dietrich College of Humanities and Social Sciences

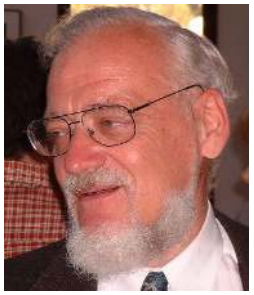


Kun Zhang

Associate Professor

Departments of Philosophy and Machine Learning

Foundations of Bayesian Epistemology



Brian Skyrms

Distinguished Professor

Department of Logic and Philosophy of Science



Louis Narens (deceased)

Professor

Mathematical Behavioral Science



Simon Huttegger

Chancellor's Professor

Department of Logic and Philosophy of Science



Teddy Seidenfeld

Herbert A. Simon University Professor of

Philosophy and Statistics

Foundations of Bayesian Epistemology





Of course, all views and errors are our own.



Simon Huttegger

CHANCELLOR'S PROFESSOR

DEPARTMENT OF LOGIC AND PHILOSOPHY OF SCIENCE



Teddy Seidenfeld

HERBERT A. SIMON UNIVERSITY PROFESSOR OF PHILOSOPHY AND STATISTICS

Aims of this talk

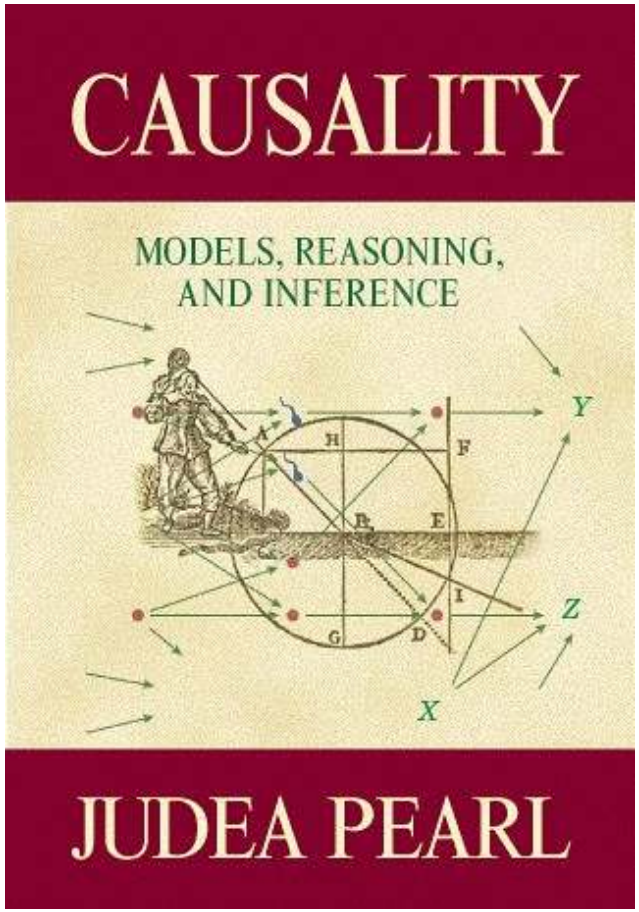
1. Argue against the (PCH) thesis—that causal reasoning is *not reducible* to probabilistic reasoning—by proving a reduction theorem in the context of structural causal models (SCMs).

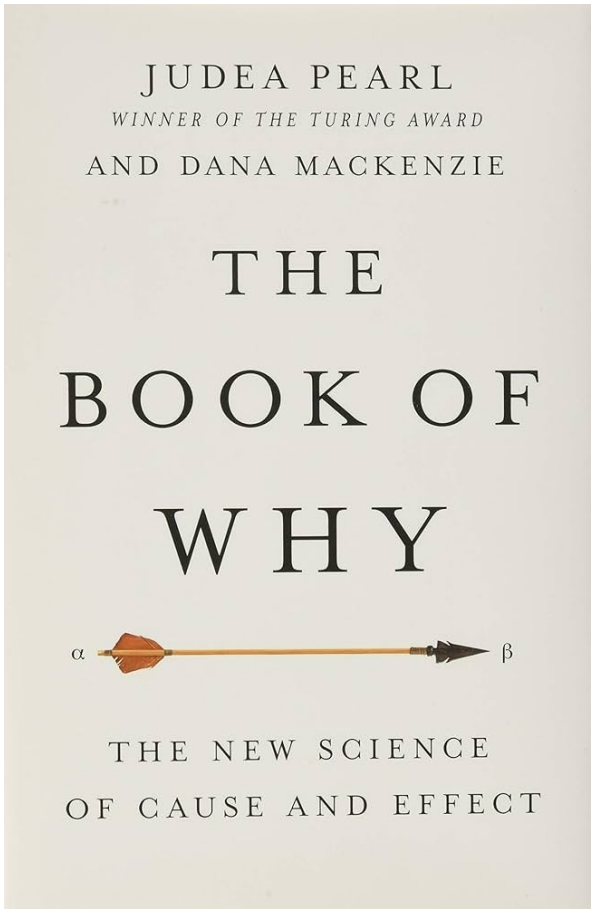
Aims of this talk

2. Advance the subjectivist project of de Finetti and friends with a representation theorem for causal learning and reasoning in terms of symmetry properties of an agent’s credences.

Aims of this talk

3. Discuss the implications of these results for the role of causation in:
    - understanding *agency*;
    - debates in *decision theory*;
    - and claims about the limits of *LLMs*.


CAUSALITY
MODELS, REASONING, AND INFERENCE
Y
Z
X
JUDEA PEARL


JUDEA PEARL
WINNER OF THE TURING AWARD
AND DANA MACKENZIE
THE
BOOK OF
WHY
α
β
THE NEW SCIENCE
OF CAUSE AND EFFECT

Pearl's Causal Hierarchy (PCH)

	Layer (Symbolic)	Typical Activity	Typical Question	Example	Machine Learning
$\mathcal{L}_3$	Counterfactual $P(y_x \mid x', y')$	Imagining	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?	
$\mathcal{L}_2$	Interventional $P(y \mid do(x), c)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured?	Reinforcement Learning
$\mathcal{L}_1$	Associational $P(y \mid x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?	Supervised/ Unsupervised Learning

“The bulk of human knowledge is organized around causal, not probabilistic relationships, and **the grammar of probability calculus is insufficient for capturing those relationships**...”

—Pearl (2001, p. 31) *Bayesianism and Causality, or, Why I am Only a Half-Bayesian*

“Specifically, the building blocks of our scientific and everyday knowledge are elementary facts such as ‘mud does not cause rain’ and ‘symptoms do not cause disease’ ...and **those facts, strangely enough, cannot be expressed in the vocabulary of probability calculus.”**

—Pearl (2001, p. 31) *Bayesianism and Causality, or, Why I am Only a Half-Bayesian*

In H. Geffner, R. Dechter, and J.Y. Halpern (Eds.), Probabilistic and Causal Inference: The Works of Judea Pearl, Association for Computing Machinery, 2021.

27

On Pearl's Hierarchy and the Foundations of Causal Inference

Elias Bareinboim (Columbia University),
Juan D. Correa (Columbia University),
Duligur Ibeling (Stanford University),
Thomas Icard (Stanford University)

Abstract

Cause-and-effect relationships play a central role in how we perceive and make sense of the world around us, how we act upon it, and ultimately, how we understand ourselves. Almost two decades ago, computer scientist Judea Pearl made a breakthrough in understanding causality by discovering and systematically studying the "Ladder of Causation," a framework that highlights the distinct roles of seeing, doing, and imagining. In honor of this landmark discovery, we name this the Pearl Causal Hierarchy (PCH). In this chapter, we develop a novel and comprehensive treatment of the PCH through two complementary lenses: one logical-probabilistic and another inferential-graphical. Following Pearl's own presentation of the hierarchy, we begin by showing how the PCH organically emerges from a well-specified collection of causal mechanisms (a structural causal model, or SCM). We then turn to the logical lens. Our first result, the Causal Hierarchy Theorem (CHT), demonstrates that the three layers of the hierarchy almost always separate in a measure-theoretic sense. Roughly speaking, the CHT says that data at one layer virtually always underdetermines information at higher layers. As in most practical settings the scientist does not have access to the precise form of the underlying causal mechanisms—only to data generated by them with respect to some of the PCH's layers—this motivates us to study inferences within the PCH through the graphical lens. Specifically, we explore a set of methods known as causal inference that enable inferences bridging the PCH's layers given a partial

Irreducibility of levels of PCH

"We prove that... it is generically impossible to draw higher-layer inferences using only lower-layer information, a result known informally in the field under the familiar adage: **'no causes-in, no causes-out.'**"

—Bareinboim et al (2021) *On Pearl's Causal Hierarchy and the Foundations of Causal Inference*, p. 513

Irreducibility of levels of PCH

"If each language did not expressively exceed its predecessors, then in some sense **our talk of causation and imagination would be no more than mere figures of speech**, being fully reducible to lower layers."

—Bareinboim et al (2021) *On Pearl's Causal Hierarchy and the Foundations of Causal Inference*, p. 527

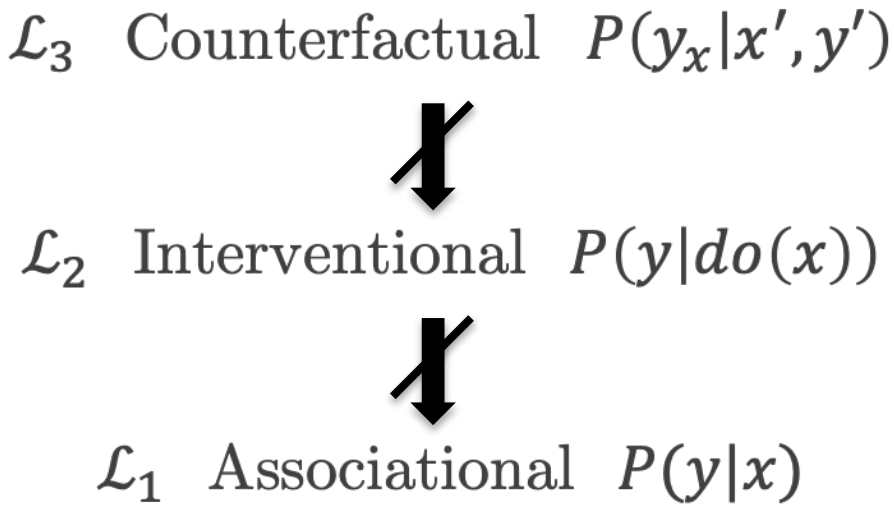
Pearl's Causal Hierarchy (PCH)

$\mathcal{L}_3$ Counterfactual $P(y_x|x', y')$

$\mathcal{L}_2$ Interventional $P(y|do(x))$

$\mathcal{L}_1$ Associational $P(y|x)$

Pearl's Causal Hierarchy (PCH)



Pearl's Causal Hierarchy (PCH)

We focus on the claim of the irreducibility of $\mathcal{L}_2$ to $\mathcal{L}_1$

$\mathcal{L}_2$ Interventional $P(y|do(x))$

$\not\downarrow$

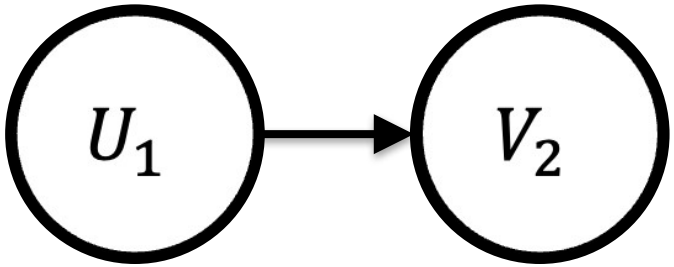
$\mathcal{L}_1$ Associational $P(y|x)$

Structural Causal Models

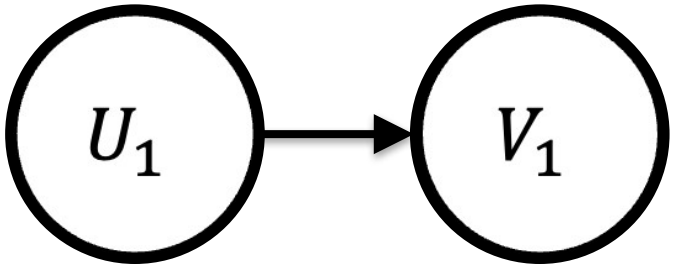
A structural causal model

$$M = \langle U, V, F, P(U) \rangle$$

Let’s consider a very simple SCM.

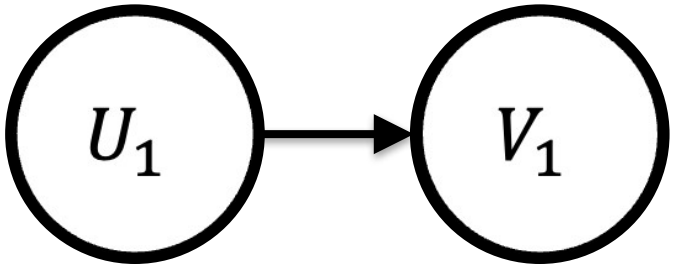

U_1
V_2

$$M = \langle U, V, F, P(U) \rangle$$



$$\mathrm{rng}(U_1) = \{0,1\}$$
$$f_{V_1} = U_1$$
$$\mathrm{rng}(V_1) = \{0,1\}$$
$$P(U_1 = 0) = P(U_1 = 1) = 0.5$$

$$M = \langle U, V, F, P(U) \rangle$$



$U = \{U_1\}$

$V = \{V_1\}$

$F = \{f_{V_1}\}$

$P(U) = \{P(U_1)\}$

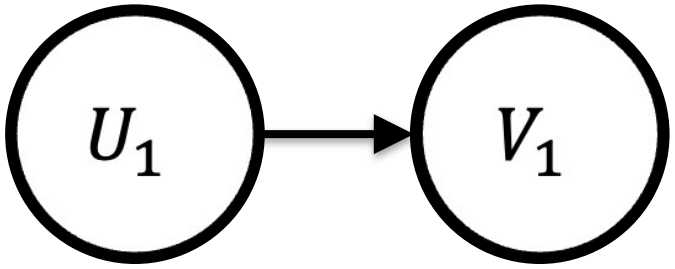
$\mathrm{rng}(U_1) = \{0,1\}$

$f_{V_1} = U_1$

$\mathrm{rng}(V_1) = \{0,1\}$

$P(U_1 = 0) = P(U_1 = 1) = 0.5$

$$M = \langle U, V, F, P(U) \rangle$$



$U = \{U_1\}$

$V = \{V_1\}$

$F = \{f_{V_1}\}$

$P(U) = \{P(U_1)\}$

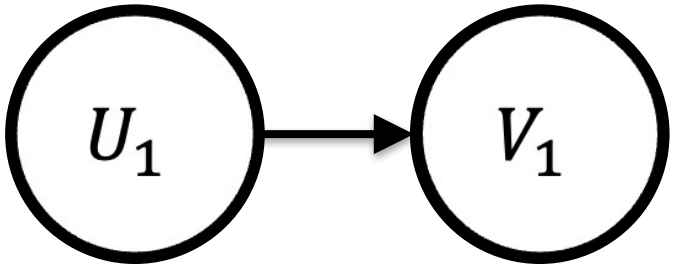
$\text{rng}(U_1) = \{0,1\}$

$f_{V_1} = U_1$

$\text{rng}(V_1) = \{0,1\}$

$P(U_1 = 0) = P(U_1 = 1) = 0.5$

$$M = \langle U, V, F, P(U) \rangle$$



$U = \{U_1\}$

$V = \{V_1\}$

$F = \{f_{V_1}\}$

$P(U) = \{P(U_1)\}$

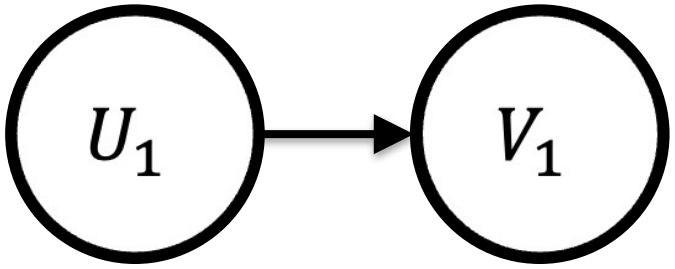
$\text{rng}(U_1) = \{0,1\}$

$f_{V_1} = U_1$

$\text{rng}(V_1) = \{0,1\}$

$P(U_1 = 0) = P(U_1 = 1) = 0.5$

$$M = \langle U, V, F, P(U) \rangle$$



$U = \{U_1\}$

$V = \{V_1\}$

$F = \{f_{V_1}\}$

$P(U) = \{P(U_1)\}$

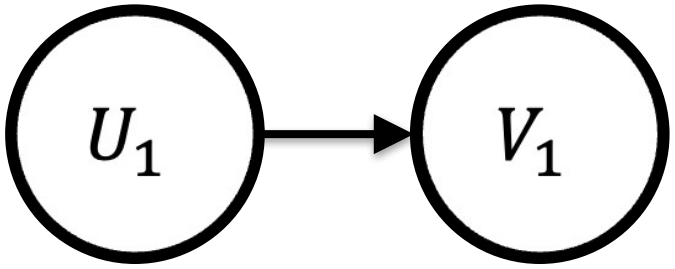
$\mathrm{rng}(U_1) = \{0,1\}$

$f_{V_1} = U_1$

$\mathrm{rng}(V_1) = \{0,1\}$

$P(U_1 = 0) = P(U_1 = 1) = 0.5$

$$M = \langle U, V, F, P(U) \rangle$$



$U = \{U_1\}$

$V = \{V_1\}$

$F = \{f_{V_1}\}$

$P(U) = \{P(U_1)\}$

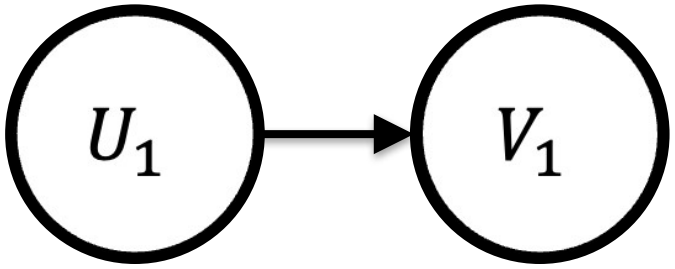
$\text{rng}(U_1) = \{0,1\}$

$f_{V_1} = U_1$

$\text{rng}(V_1) = \{0,1\}$

$P(U_1 = 0) = P(U_1 = 1) = 0.5$

$$M = \langle U, V, F, P(U) \rangle$$



$U = \{U_1\}$

$V = \{V_1\}$

$F = \{f_{V_1}\}$

$P(U) = \{P(U_1)\}$

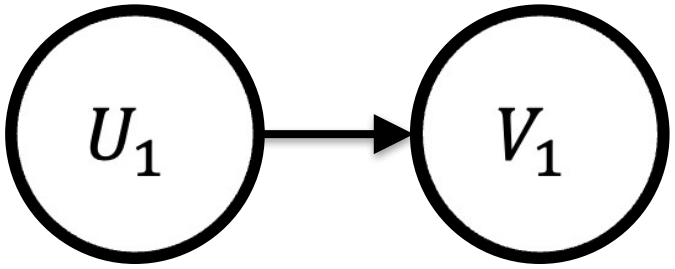
$\text{rng}(U_1) = \{0,1\}$

$f_{V_1} = U_1$

$\text{rng}(V_1) = \{0,1\}$

$P(U_1 = 0) = P(U_1 = 1) = 0.5$

$$M = \langle U, V, F, P(U) \rangle$$



$U = \{U_1\}$

$V = \{V_1\}$

$F = \{f_{V_1}\}$

$P(U) = \{P(U_1)\}$

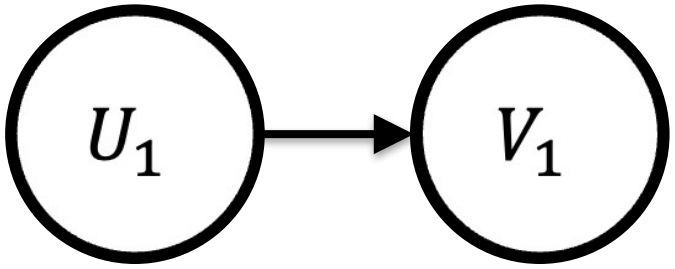
$\text{rng}(U_1) = \{0,1\}$

$f_{V_1} = U_1$

$\text{rng}(V_1) = \{0,1\}$

$P(U_1 = 0) = P(U_1 = 1) = 0.5$

$$M = \langle U, V, F, P(U) \rangle$$



$U = \{U_1\}$

$V = \{V_1\}$

$F = \{f_{V_1}\}$

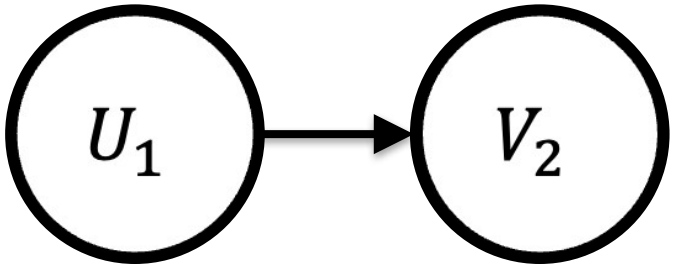
$P(U) = \{P(U_1)\}$

$\text{rng}(U_1) = \{0,1\}$

$f_{V_1} = U_1$

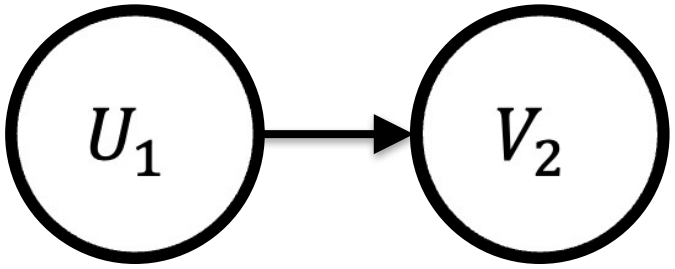
$\text{rng}(V_1) = \{0,1\}$

$P(U_1 = 0) = P(U_1 = 1) = 0.5$



$$\mathcal{L}_1 : P(X = x | Y = y) \text{ for } X, Y \in \{U_1, V_1\}$$

$$\mathcal{L}_2 : P(X = x | \text{do}(Y = y)) \text{ for } X, Y \in \{U_1, V_1\}$$



$$\mathcal{L}_1 : P(X = x | Y = y) \text{ for } X, Y \in \{U_1, V_1\}$$

$$\mathcal{L}_2 : P(X = x | \text{do}(Y = y)) \text{ for } X, Y \in \{U_1, V_1\}$$

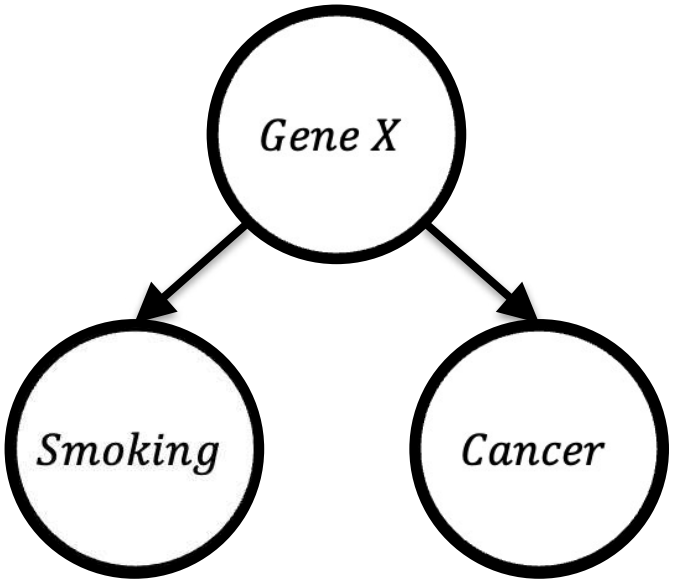
We consider SCMs exclusively in their capacity as representing *type-level* causal relations.

For now, we restrict our focus to models satisfying:

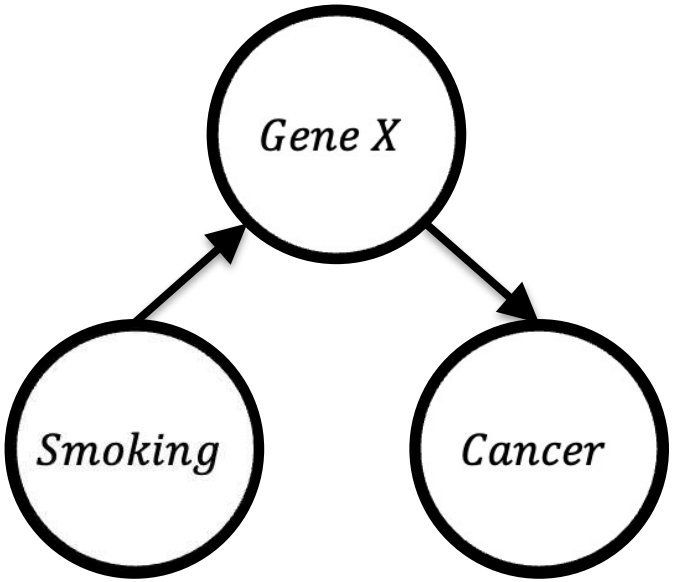
1. Causal Markov Condition

2. Faithfulness

Example of irreducibility of $\mathcal{L}_2$ to $\mathcal{L}_1$

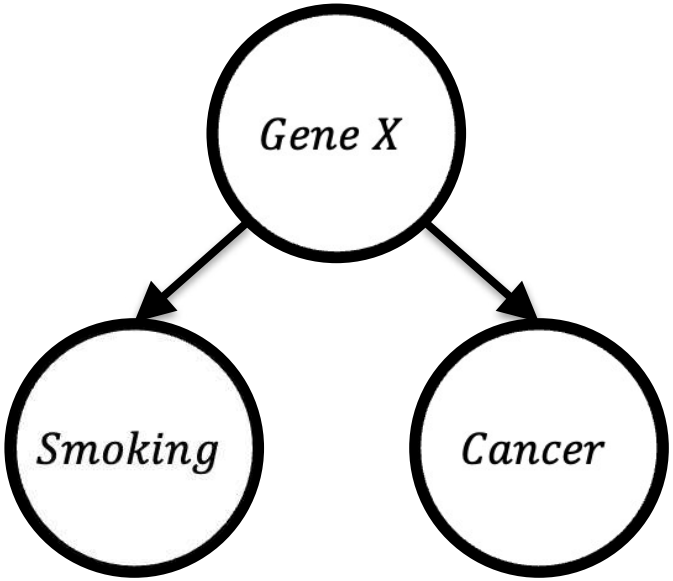


M_1: $Smoking \perp Cancer \mid Gene\ X$

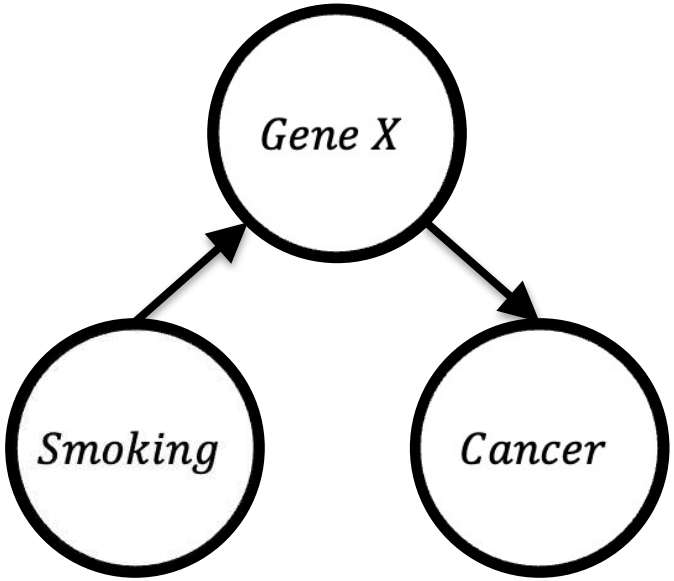


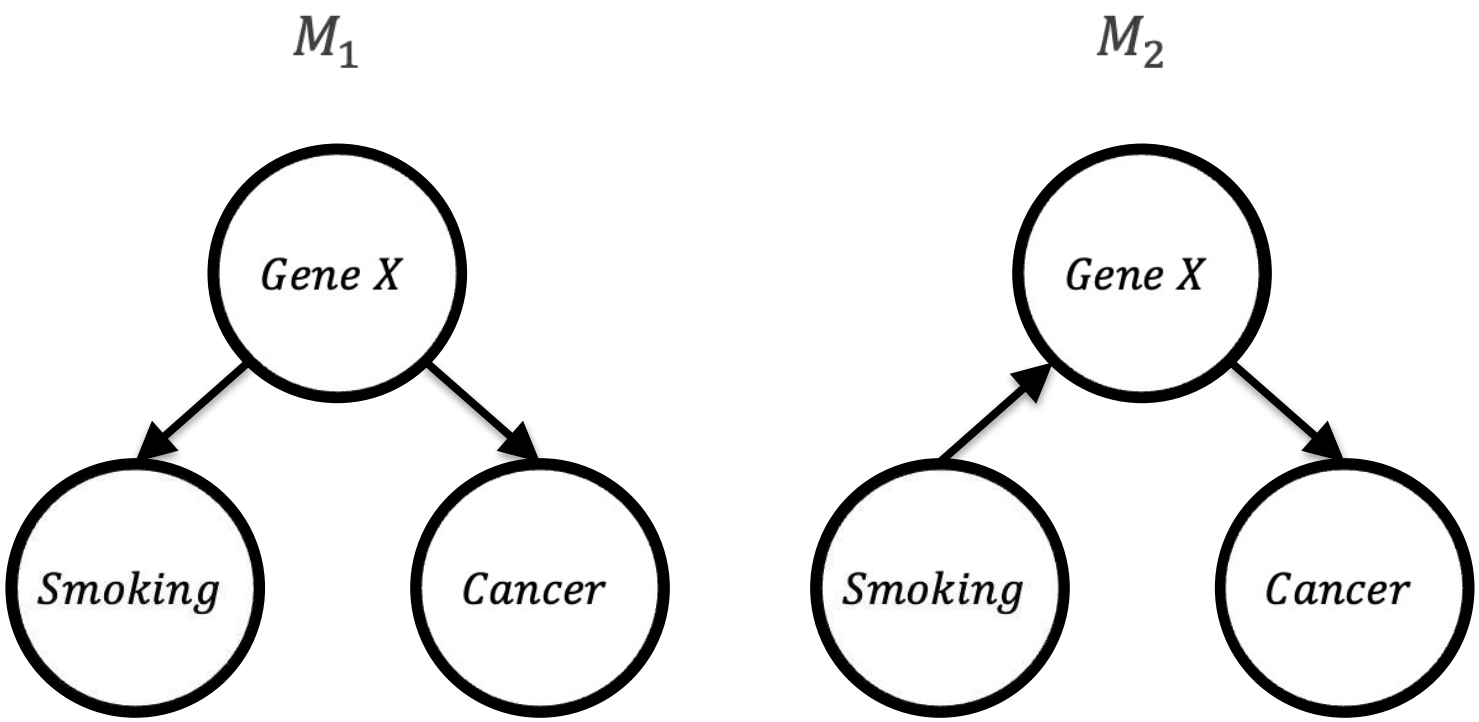
M_2: $Smoking \perp Cancer \mid Gene\ X$

M_1

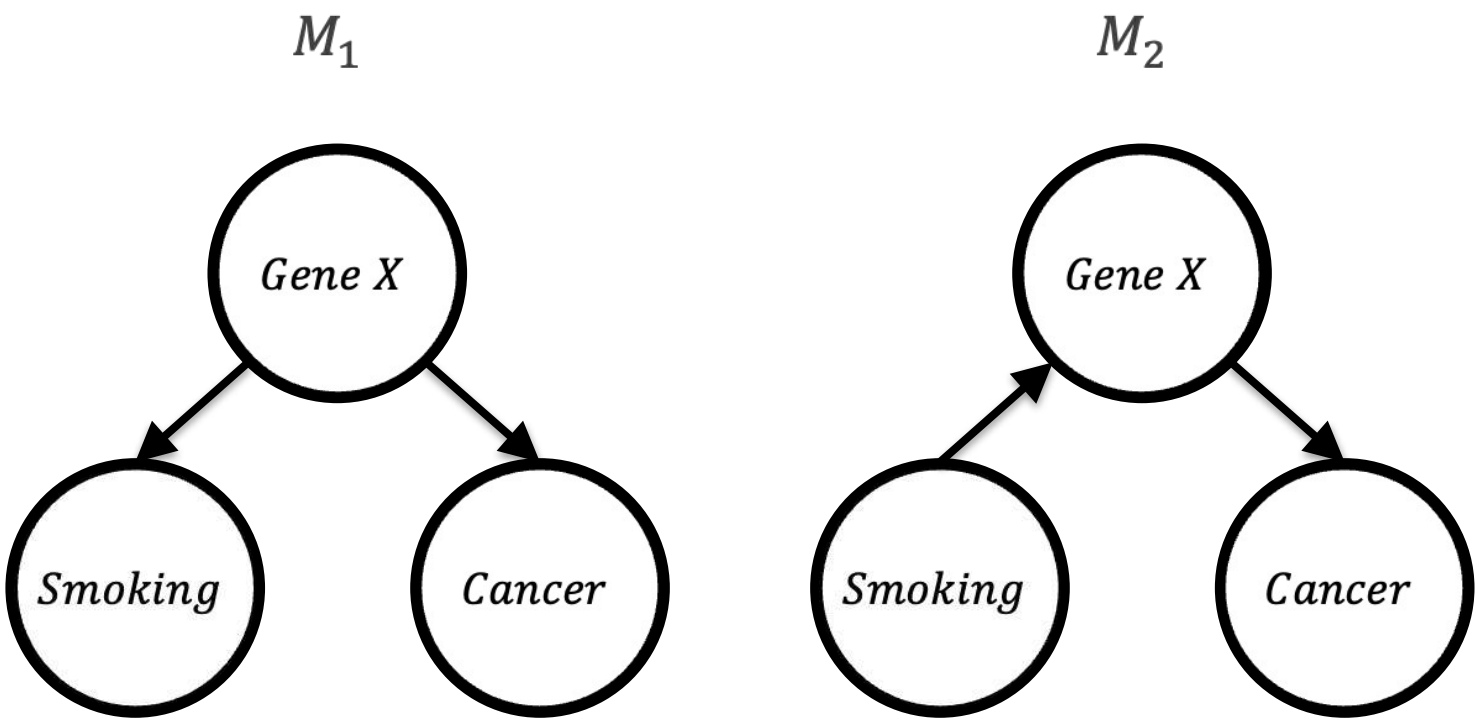


M_2

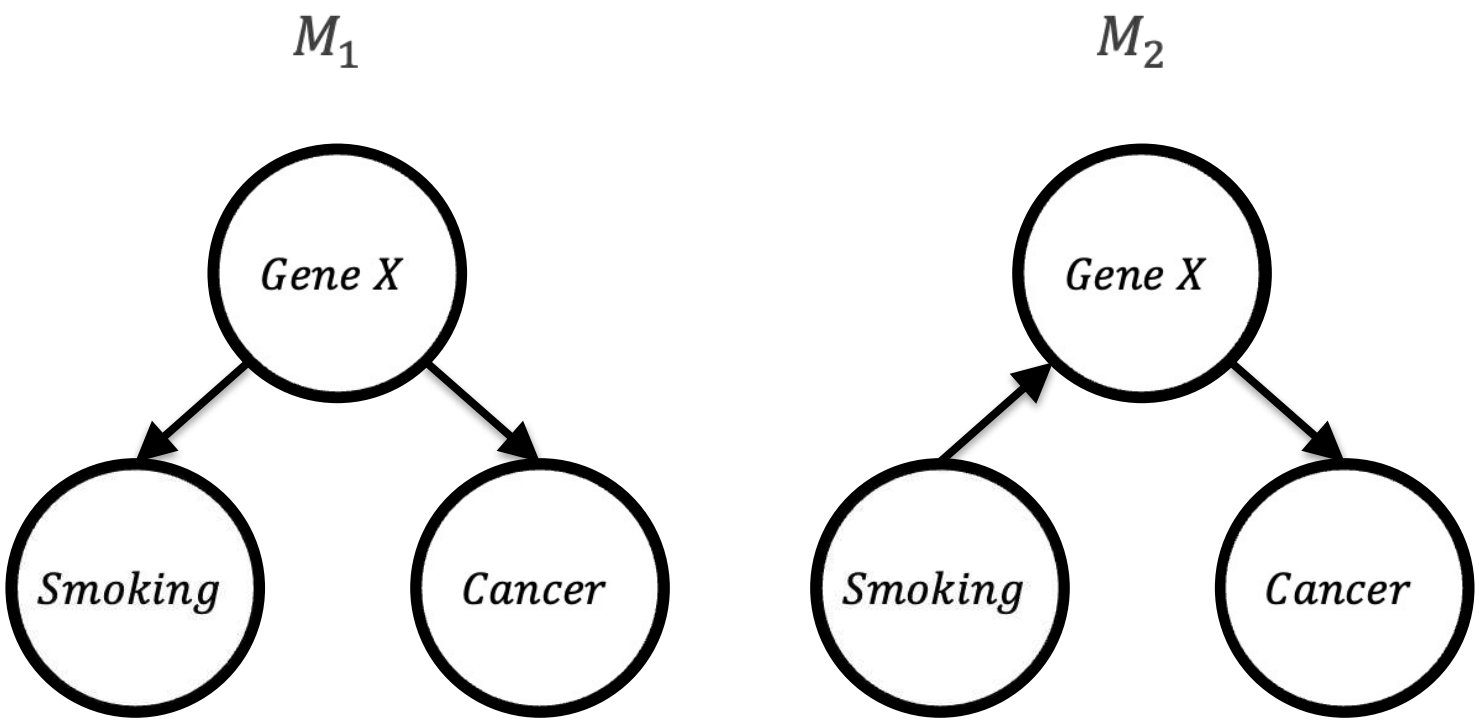




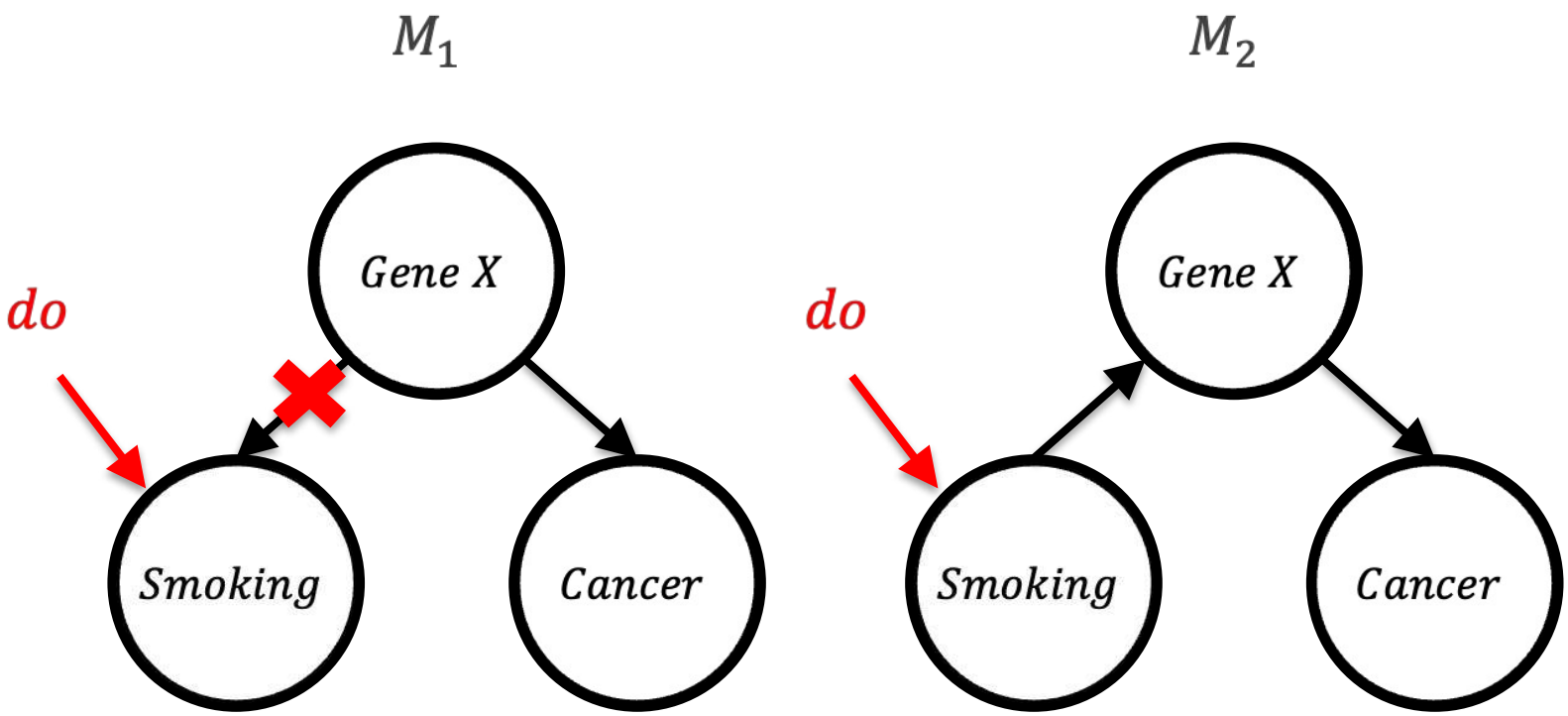
Fact. These two graphs are *indistinguishable* using only observations of realizations of *these variables*.



$$P_{M_1}\ (Cancer|Smoking) = P_{M_2}\ (Cancer|Smoking)$$



However, they *can be distinguished* if we have access to observations of interventions.

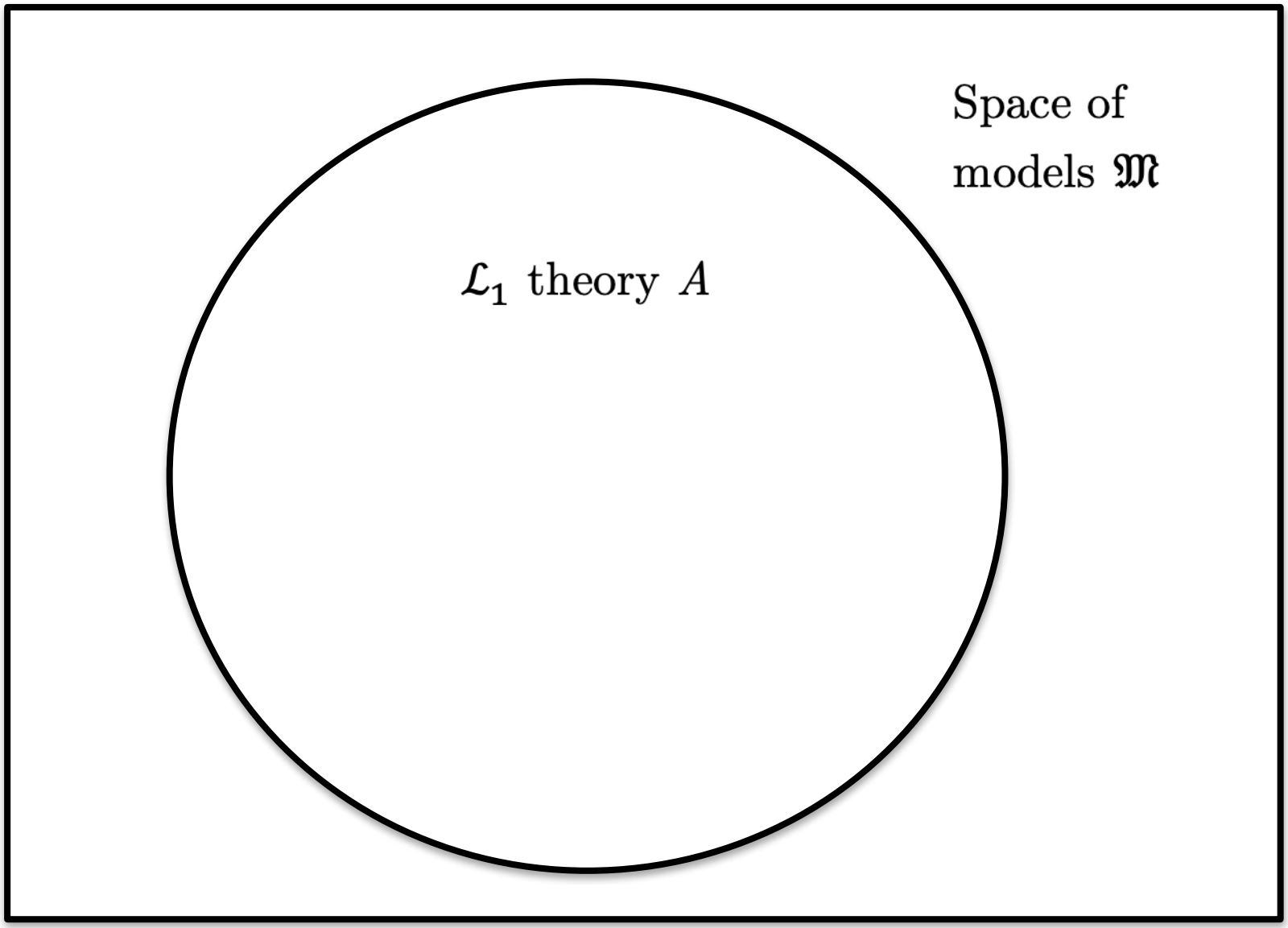


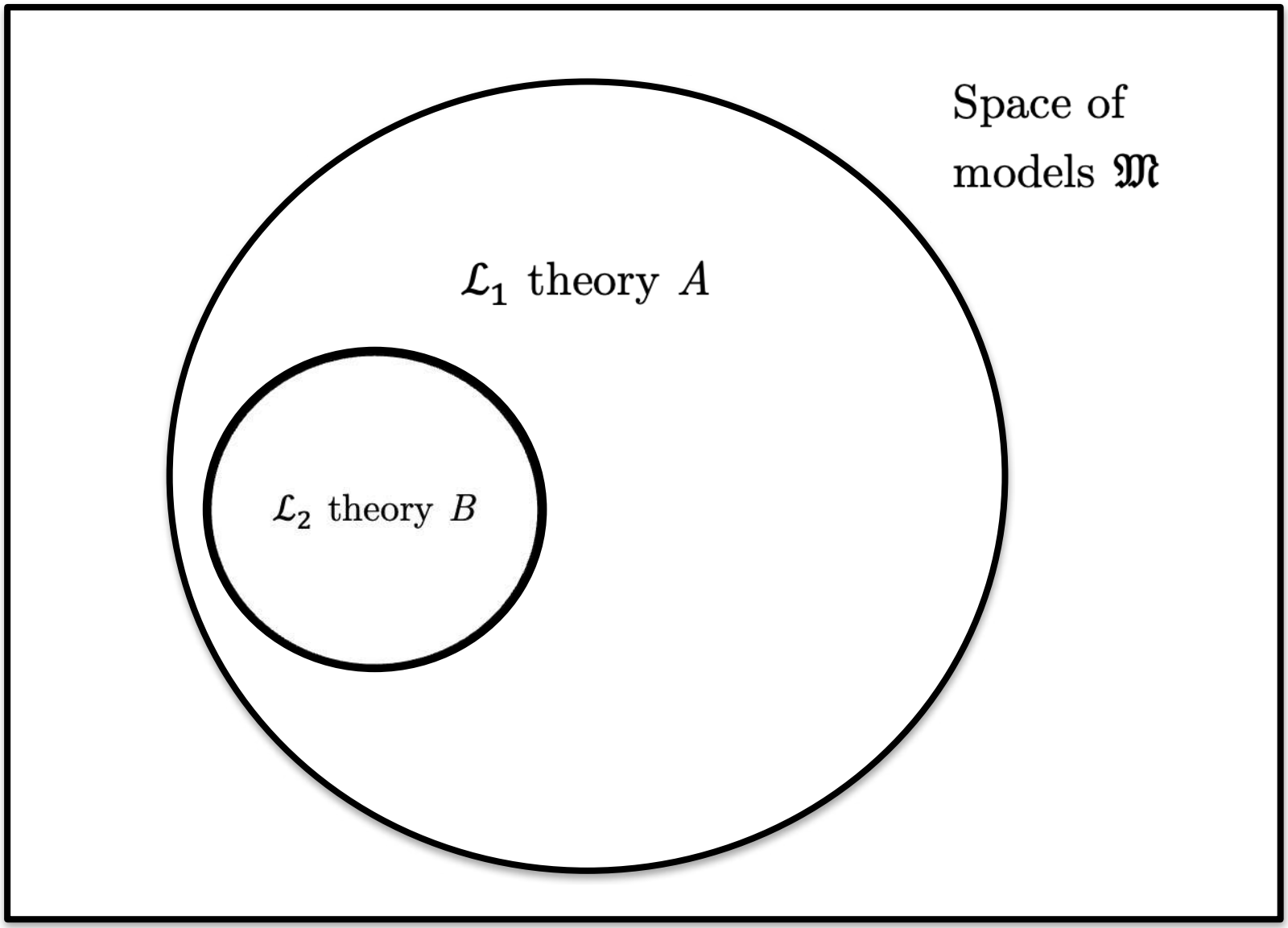
$$P_{M_1}(Cancer|do(Smoking)) \neq P_{M_2}(Cancer|do(Smoking))$$

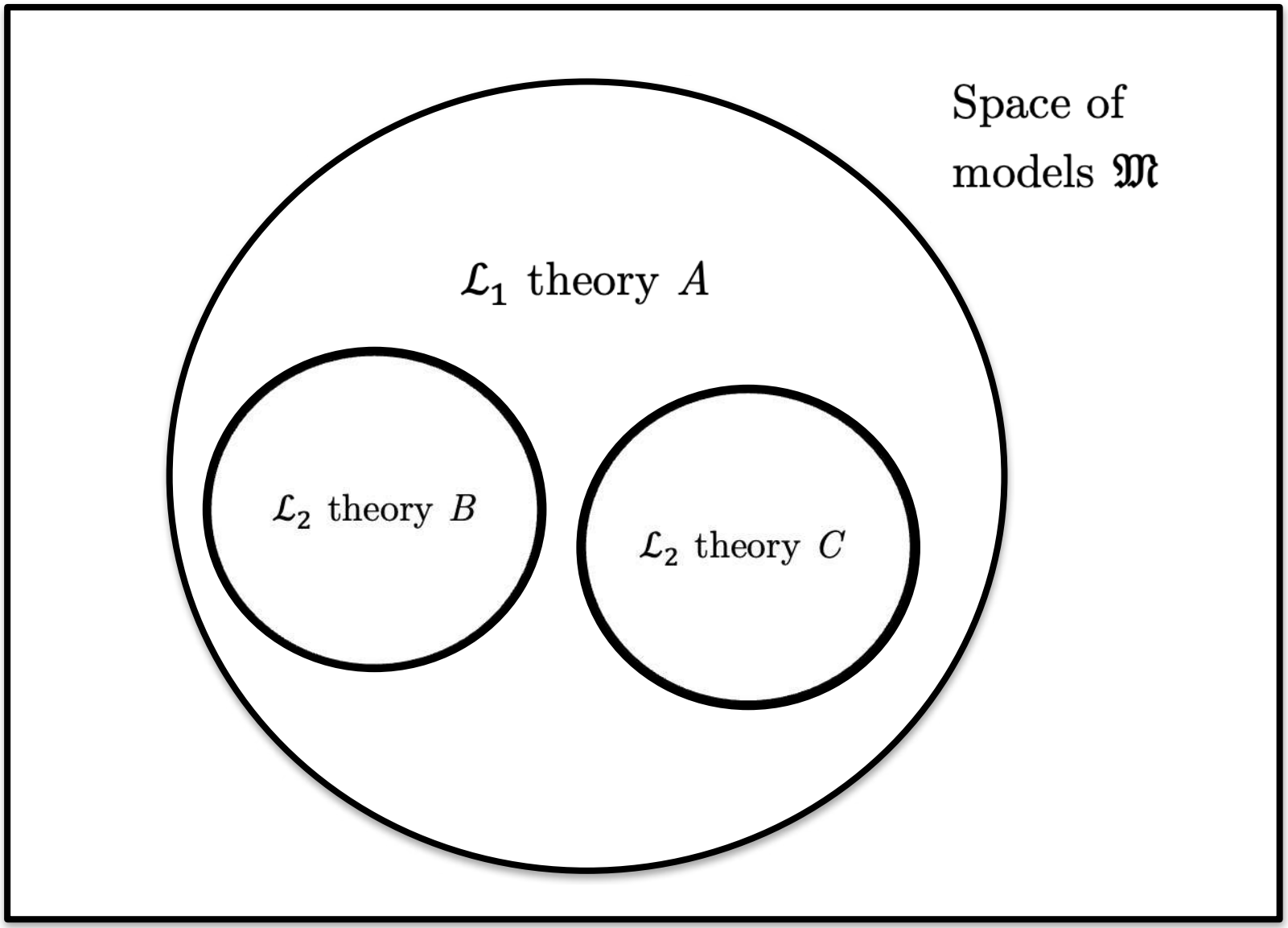
Causal Hierarchy Theorem (CHT)

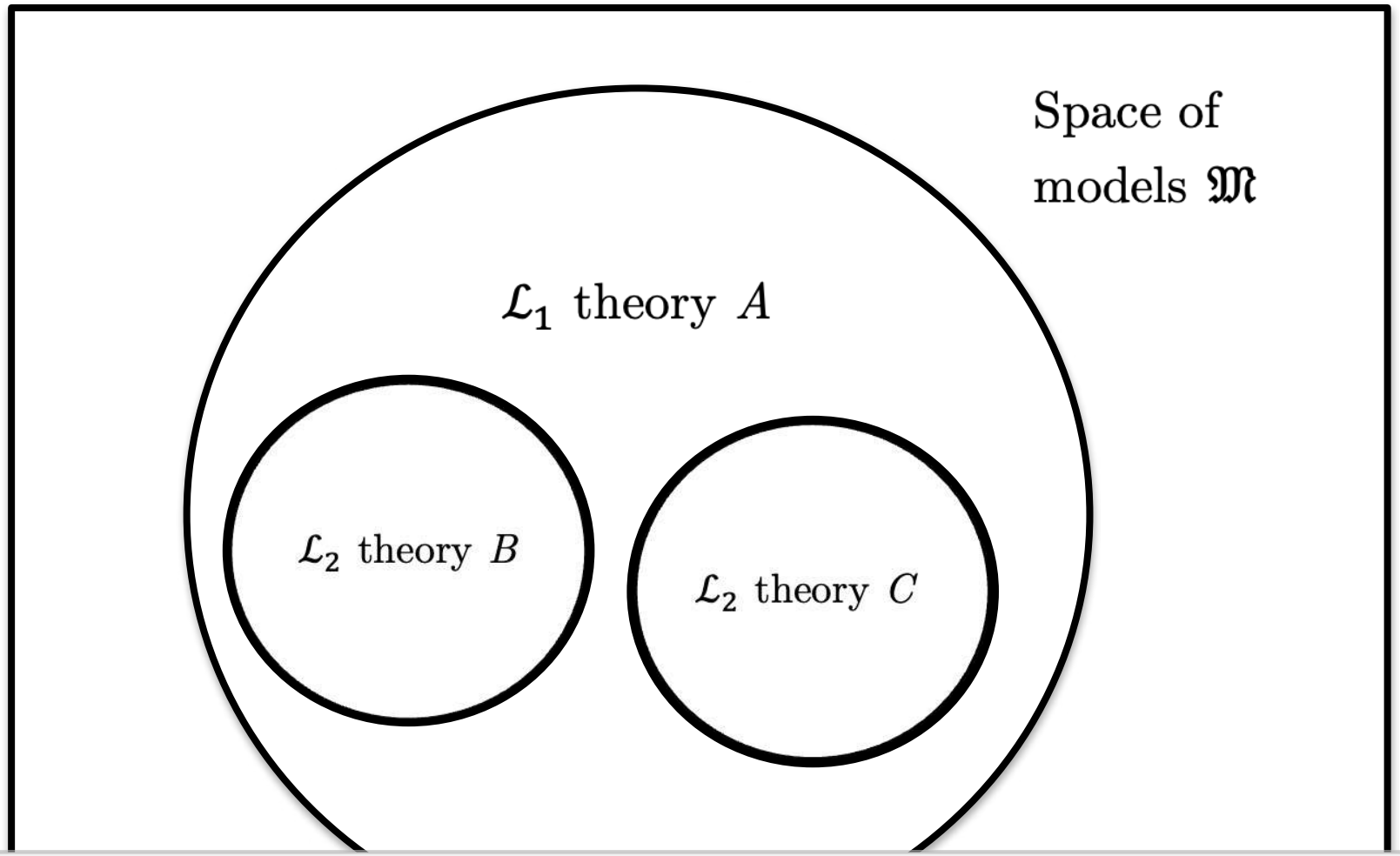
Theorem. The PCH almost never collapses. That is, for almost any SCM, the layers of the hierarchy remain distinct.

Space of
models $\mathfrak{M}$

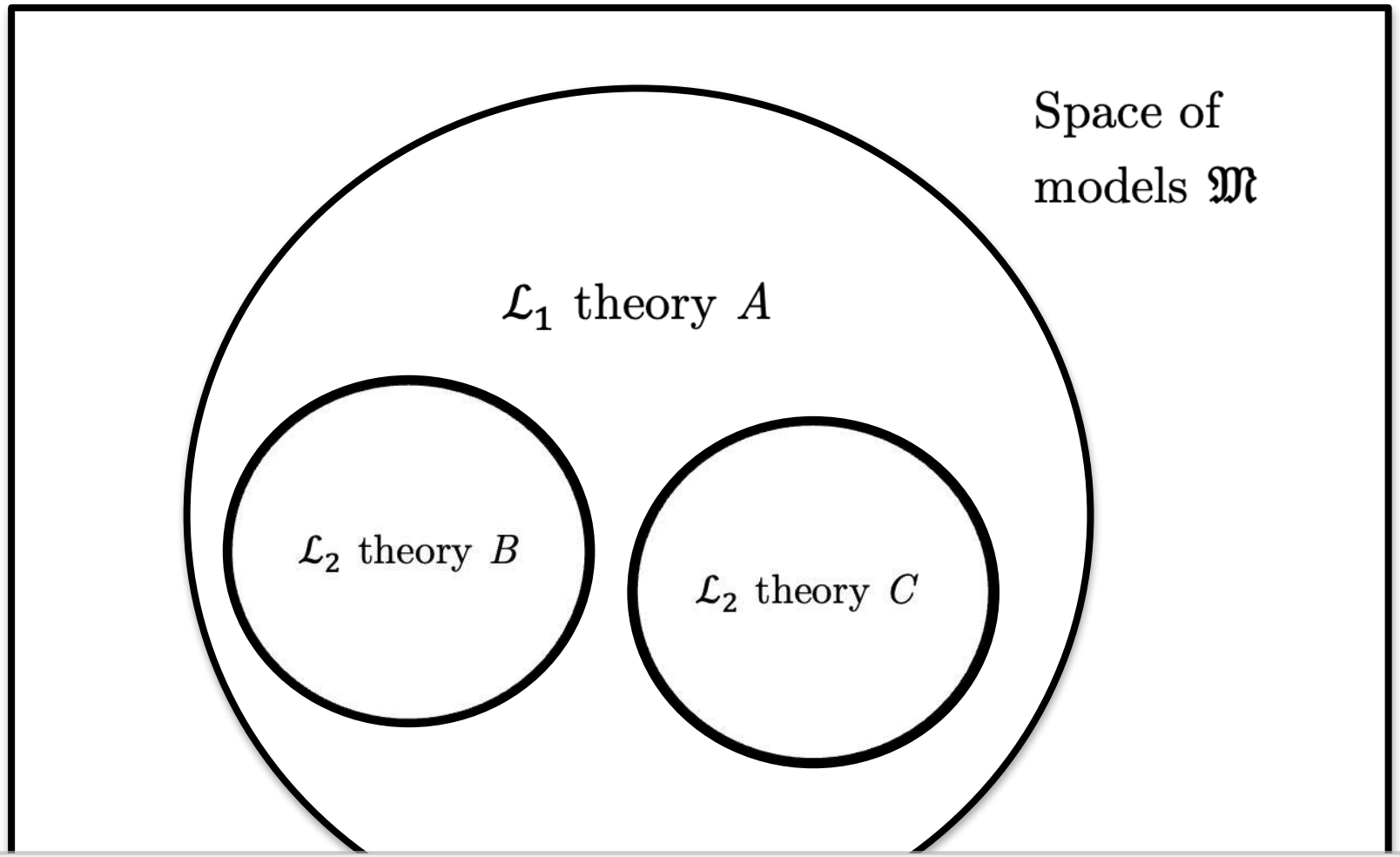

Space of
models $\mathfrak{M}$
$\mathcal{L}_1$ theory A


Space of
models $\mathfrak{M}$
$\mathcal{L}_1$ theory A
$\mathcal{L}_2$ theory B


Space of
models $\mathfrak{M}$
$\mathcal{L}_1$ theory A
$\mathcal{L}_2$ theory B
$\mathcal{L}_2$ theory C



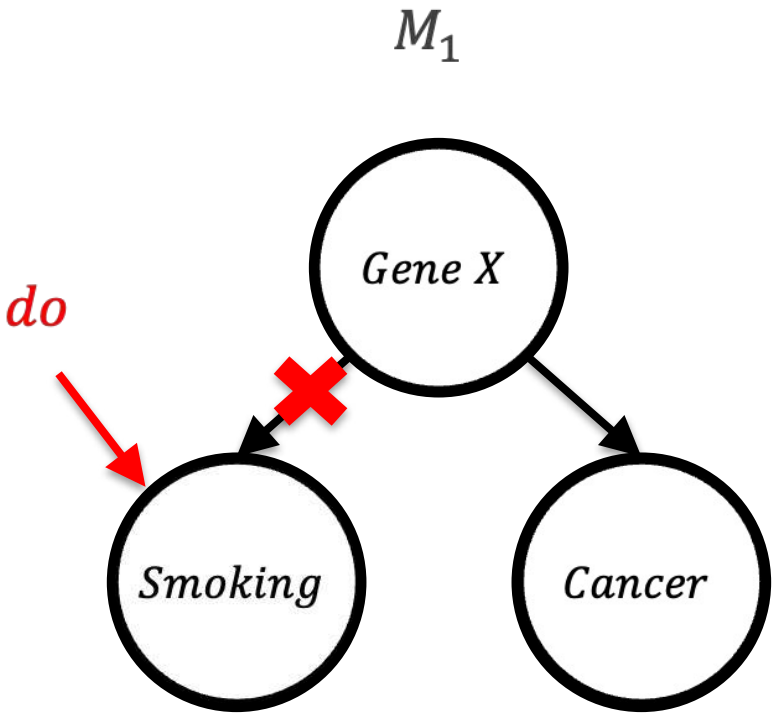
An $\mathcal{L}_1$ theory is compatible with multiple $\mathcal{L}_2$ theories.

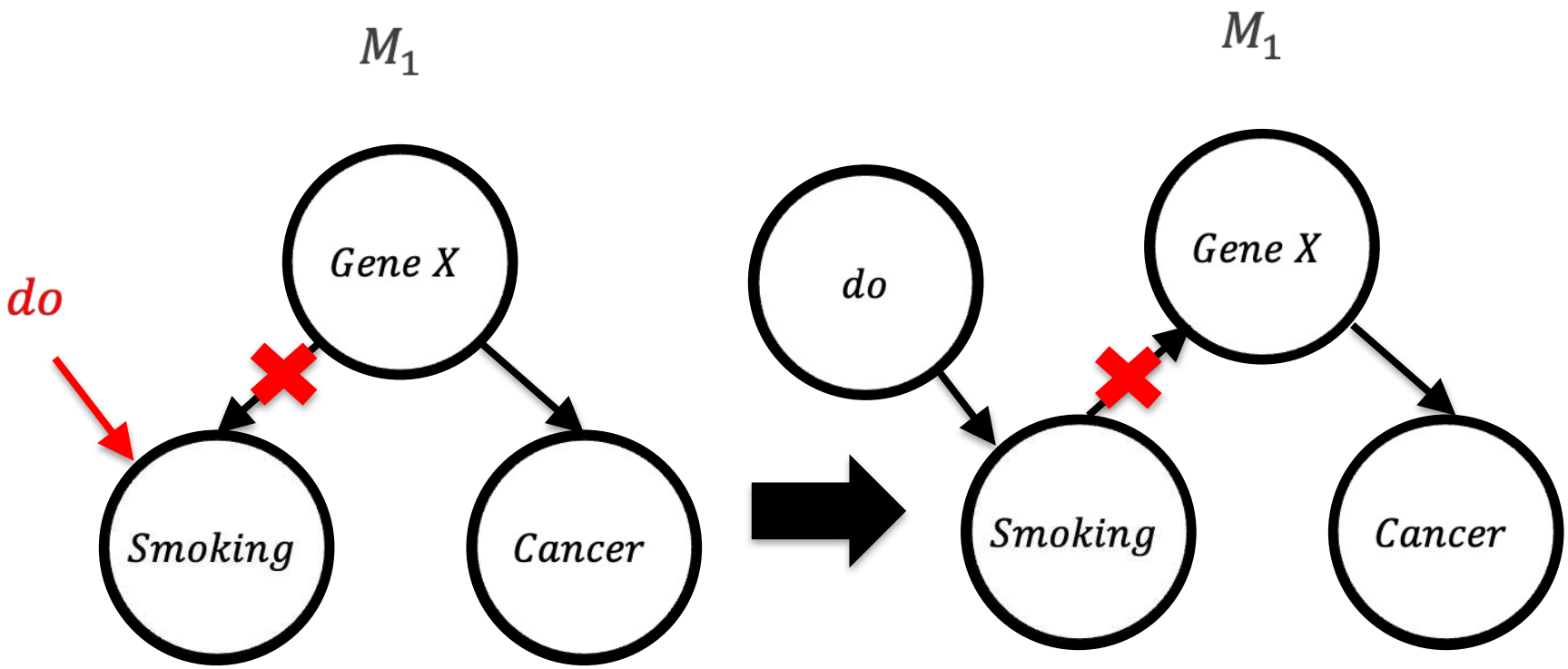


Given a natural measure over $\mathfrak{M}$, this is generically true.

The proof is valid. But we think that a straightforwardly enriched associational language can reproduce the interventional semantics.

N.B.: Interventions encoded by do-operations that are made *exogenous* to one graph can be made *endogenous* to another graph.


M_1
Gene X
do
Smoking
Cancer


M_1
Gene X
do
Smoking
Cancer
M_1
do
Gene X
Smoking
Cancer

Given this insight, can we prove the reducibility of interventional claims to associational ones?

Lecture Notes in Statistics 81

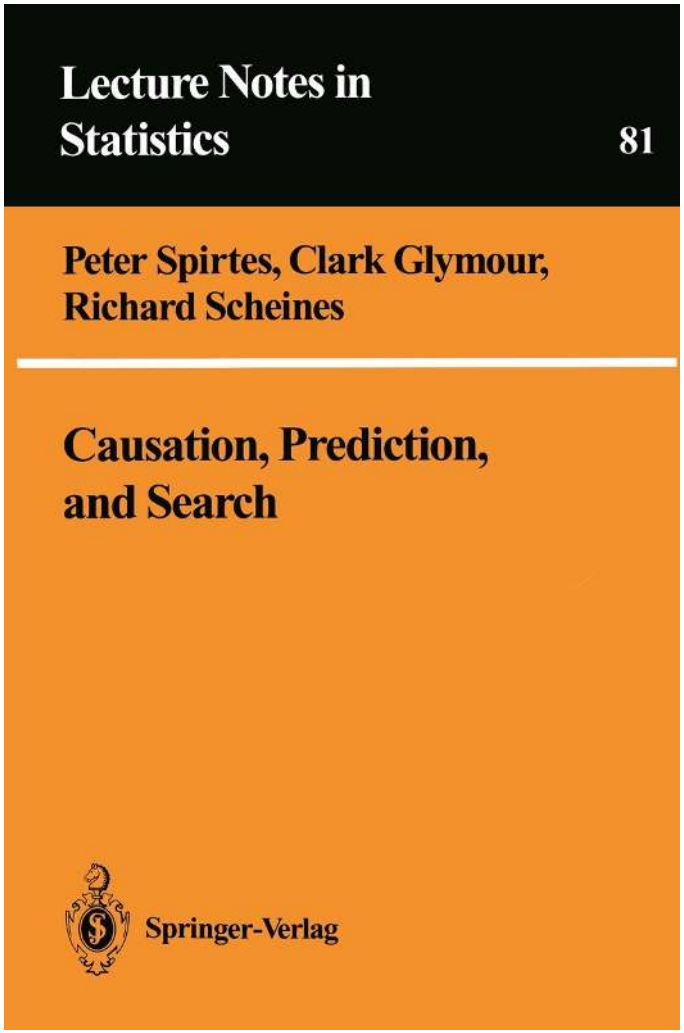
Peter Spirtes, Clark Glymour,
Richard Scheines

Causation, Prediction, and Search

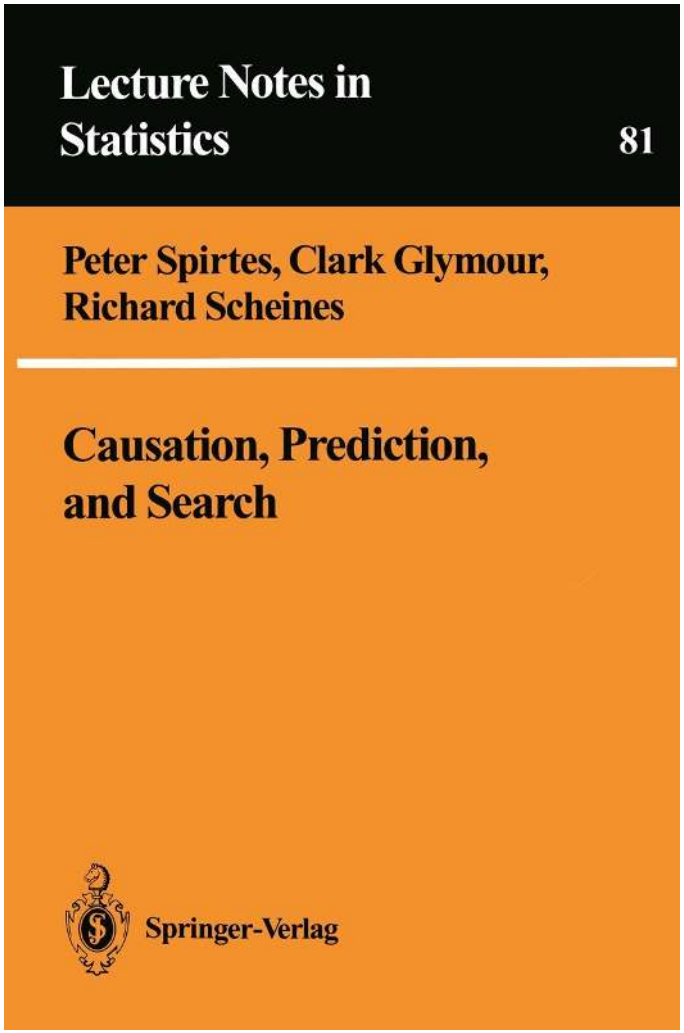


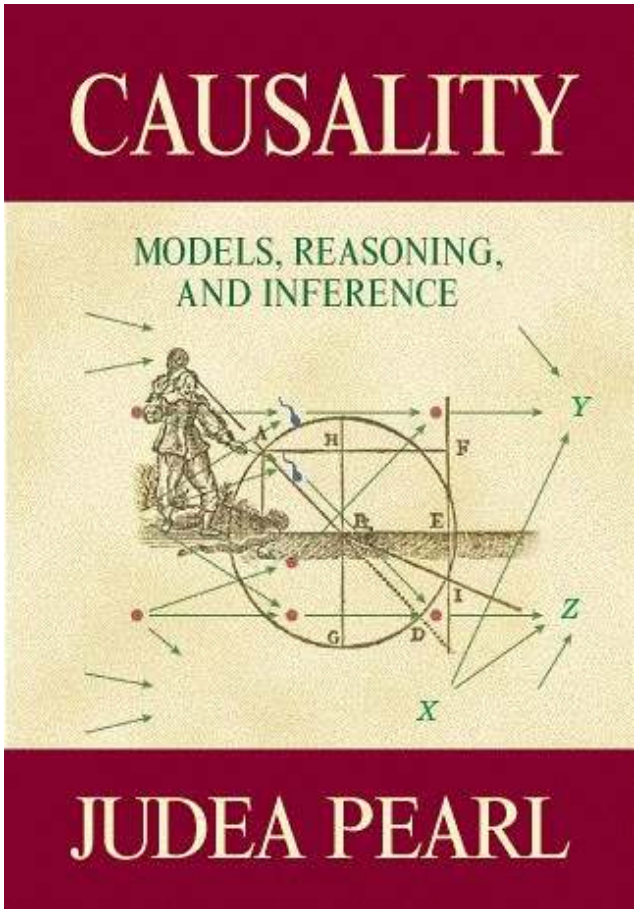
Springer-Verlag

Spirtes et al (1993) prove (Ch.4, Thm 4.6) that any two DAGs can be distinguished **provided additional variables** with the appropriate causal structure exist and can be measured.



That is, any DAG is uniquely identifiable from observational data given the right sorts of additional variables.





Even Pearl (2000, p. 99) observes that the do-operator can be emulated by adjoining a new exogenous variable I_X that fixes the value of X.

- *A Reduction Theorem for $\mathcal{L}_2$ to $\mathcal{L}_1$*

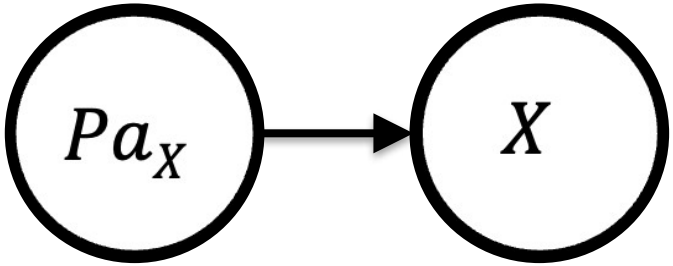
Reduction Theorem

Theorem 1. For any structural causal model, M, one can produce an extension, M' of M, such that all the interventional propositions true of M are expressible in terms of associational propositions true of M' and all the associational proposition of M continue to be true of M'.

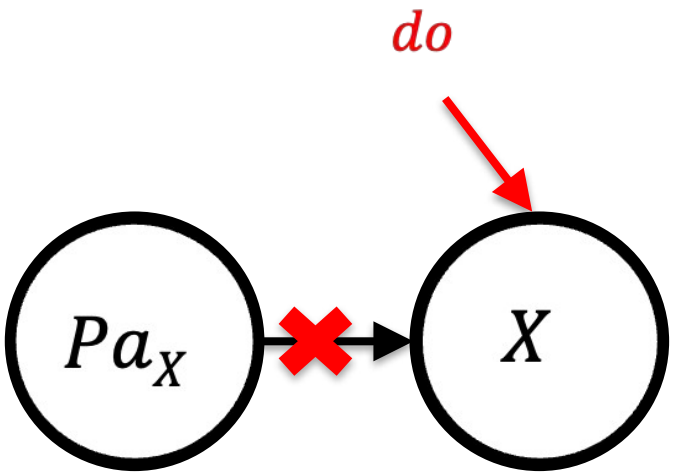
For any SCM, we can construct a conservative extension by introducing **auxiliary variables** and **modified structural equations** that replicate the effects of interventions.

This allows the use of **conditioning** in place of *do*-operations while preserving the content of both associational and interventional propositions from the original SCM.

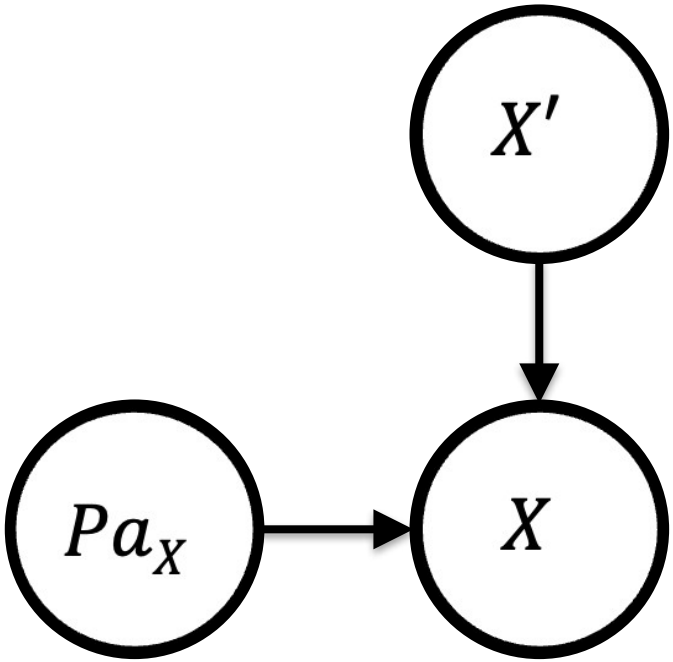
Old M



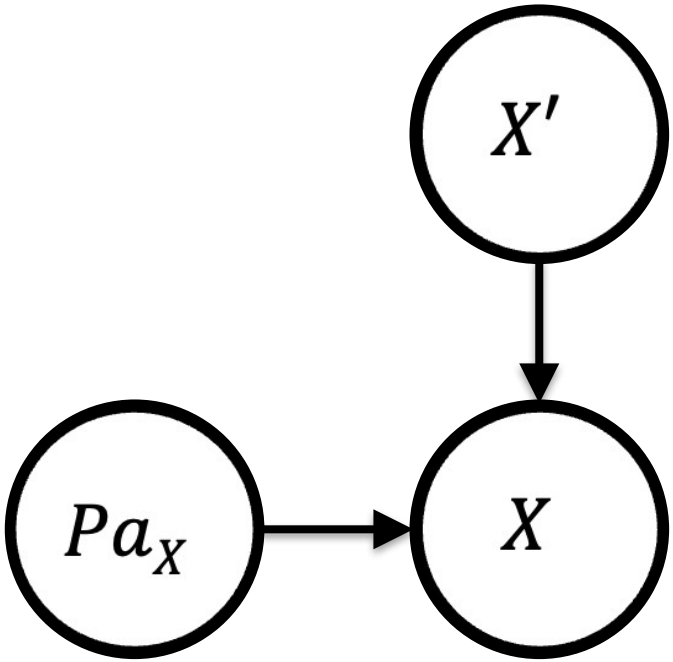
Old M



New M



New M

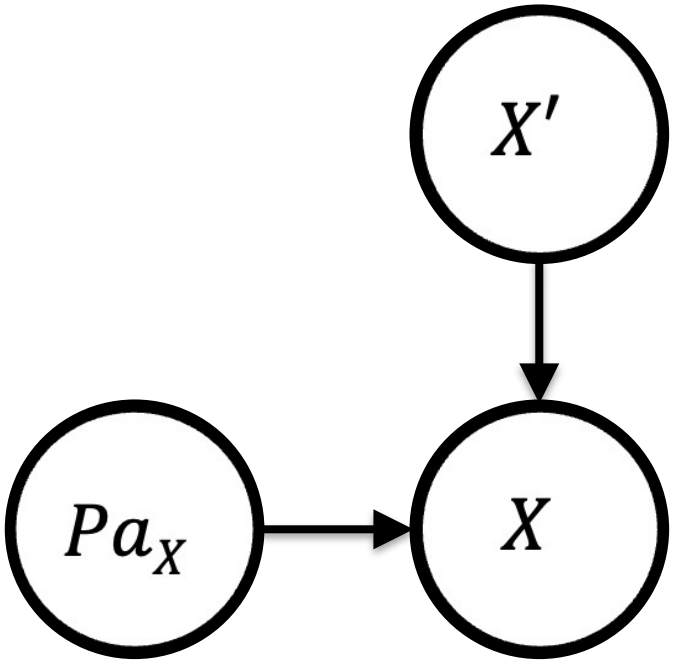


$$Rng(X') = Rng(X) \cup \{\text{NA}\}$$

NA indicates no intervention is performed on X.

$$f_{X_{\text{new}}} = \begin{cases} X', & X' \neq \text{NA} \\ f_{X_{\text{old}}}, & X' = \text{NA} \end{cases}$$

New M

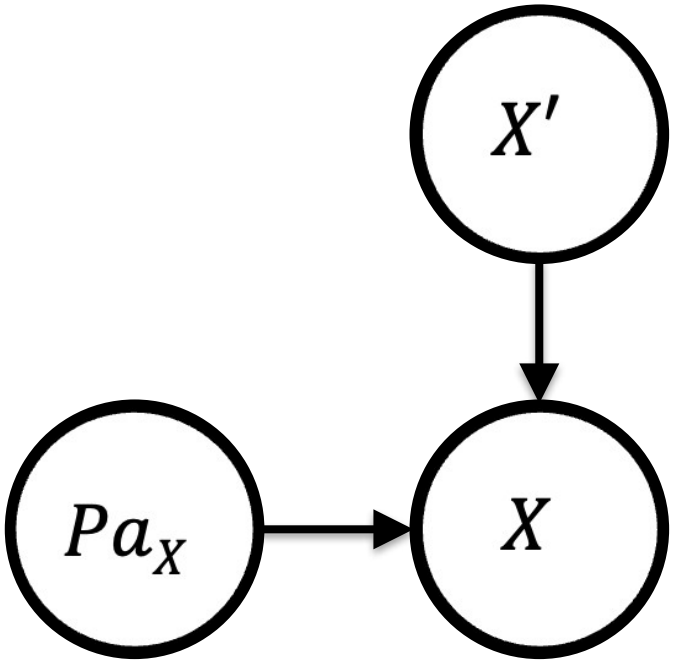


$Rng(X') = Rng(X) \cup \{\text{NA}\}$

NA indicates no intervention is performed on X.

$$f_{X_{\text{new}}} = \begin{cases} X', & X' \neq \text{NA} \\ f_{X_{\text{old}}}, & X' = \text{NA} \end{cases}$$

New M



$Rng(X') = Rng(X) \cup \{\text{NA}\}$

NA indicates no intervention is performed on X.

$$f_{X_{\text{new}}} = \begin{cases} X', & X' \neq \text{NA} \\ f_{X_{\text{old}}}, & X' = \text{NA} \end{cases}$$

In places, Pearl articulates a view not dissimilar to ours...

“If you wish to include the entire universe in the model, **causality disappears because interventions disappear**—the manipulator and the manipulated lose their distinction.

“If you wish to include the entire universe in the model, **causality disappears because interventions disappear**—the manipulator and the manipulated lose their distinction. However, scientists rarely consider the entirety of the universe as an object of investigation.

“If you wish to include the entire universe in the model, **causality disappears because interventions disappear**—the manipulator and the manipulated lose their distinction. However, scientists rarely consider the entirety of the universe as an object of investigation. In most cases the scientist carves a piece from the universe and proclaims that piece in—namely, the focus of investigation.

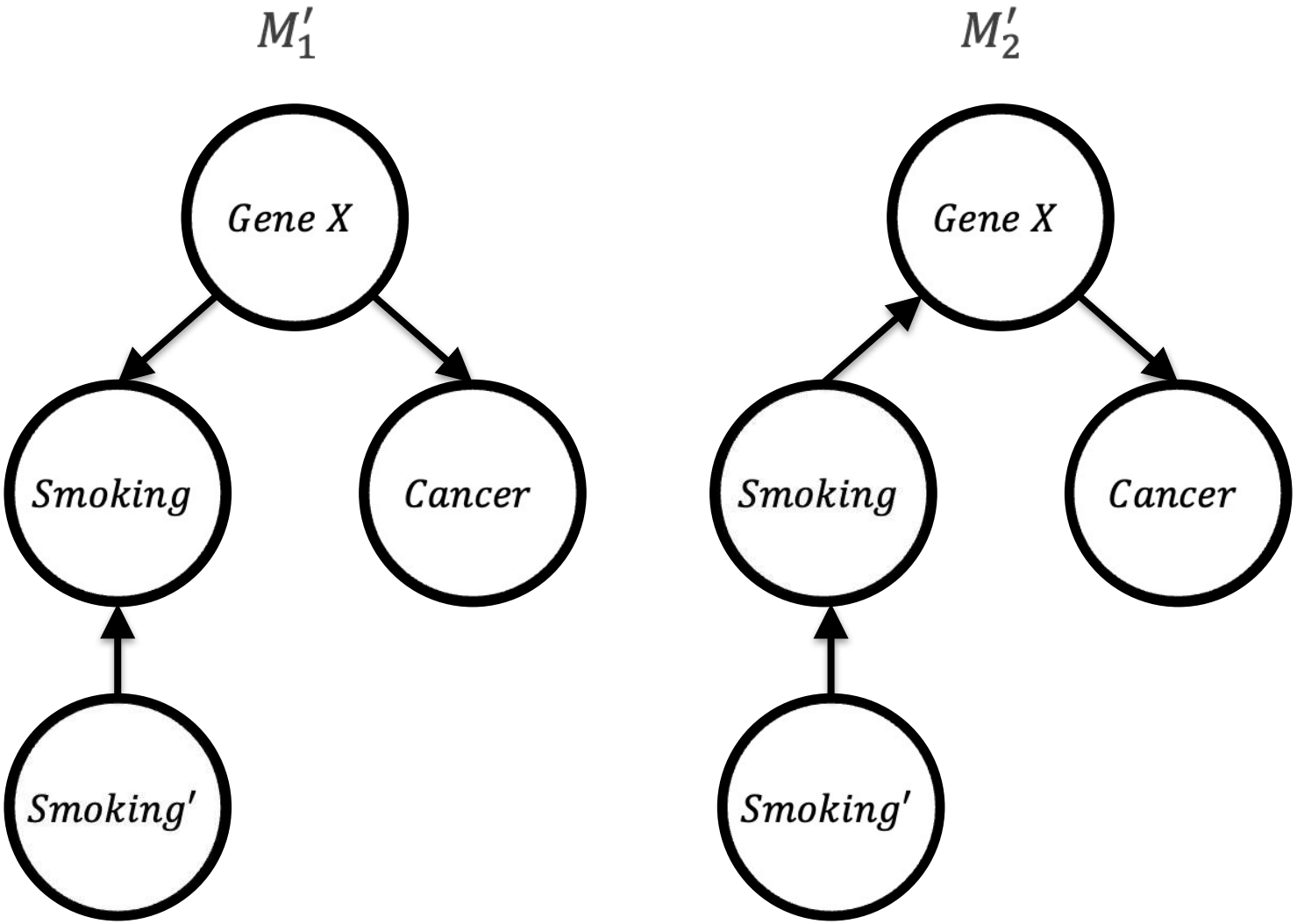
“If you wish to include the entire universe in the model, **causality disappears because interventions disappear**—the manipulator and the manipulated lose their distinction. However, scientists rarely consider the entirety of the universe as an object of investigation. In most cases the scientist carves a piece from the universe and proclaims that piece in—namely, the focus of investigation. The rest of the universe is then considered out or background and is summarized by what we call boundary conditions.

“If you wish to include the entire universe in the model, **causality disappears because interventions disappear**—the manipulator and the manipulated lose their distinction. However, scientists rarely consider the entirety of the universe as an object of investigation. In most cases the scientist carves a piece from the universe and proclaims that piece in—namely, the focus of investigation. The rest of the universe is then considered out or background and is summarized by what we call boundary conditions. This choice of ins and outs creates asymmetry in the way we look at things, and it is this asymmetry that permits us to talk about ‘outside intervention’ and hence about causality and cause-effect directionality.”

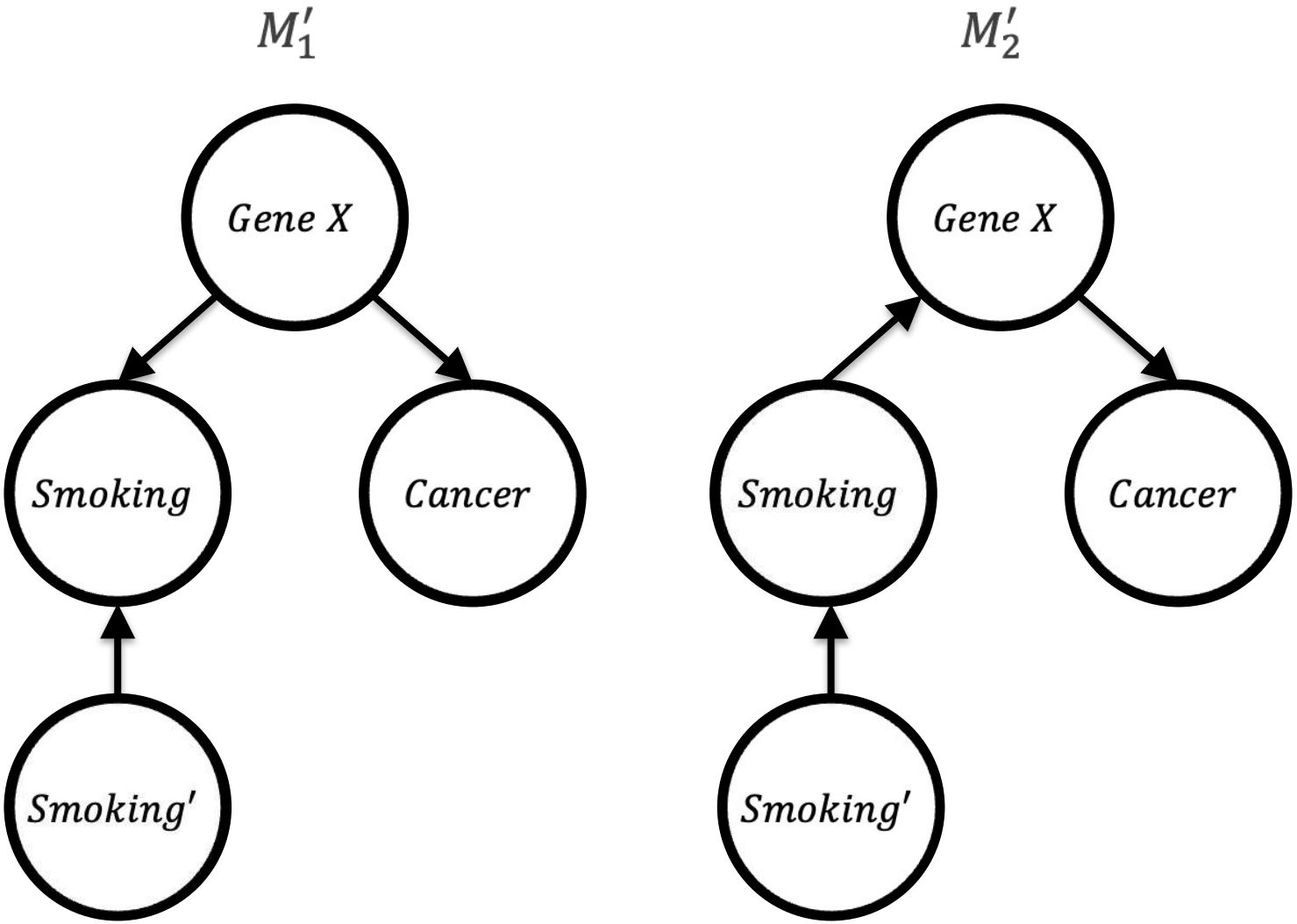
—Judea Pearl (p. 419f. of the Epilogue of the 2nd ed of *Causality*)

Note: One upshot of this picture is that causal reasoning may help us precisely when we have a smaller conceptual space—this helps explains its value and prevalence as a pattern of reasoning.

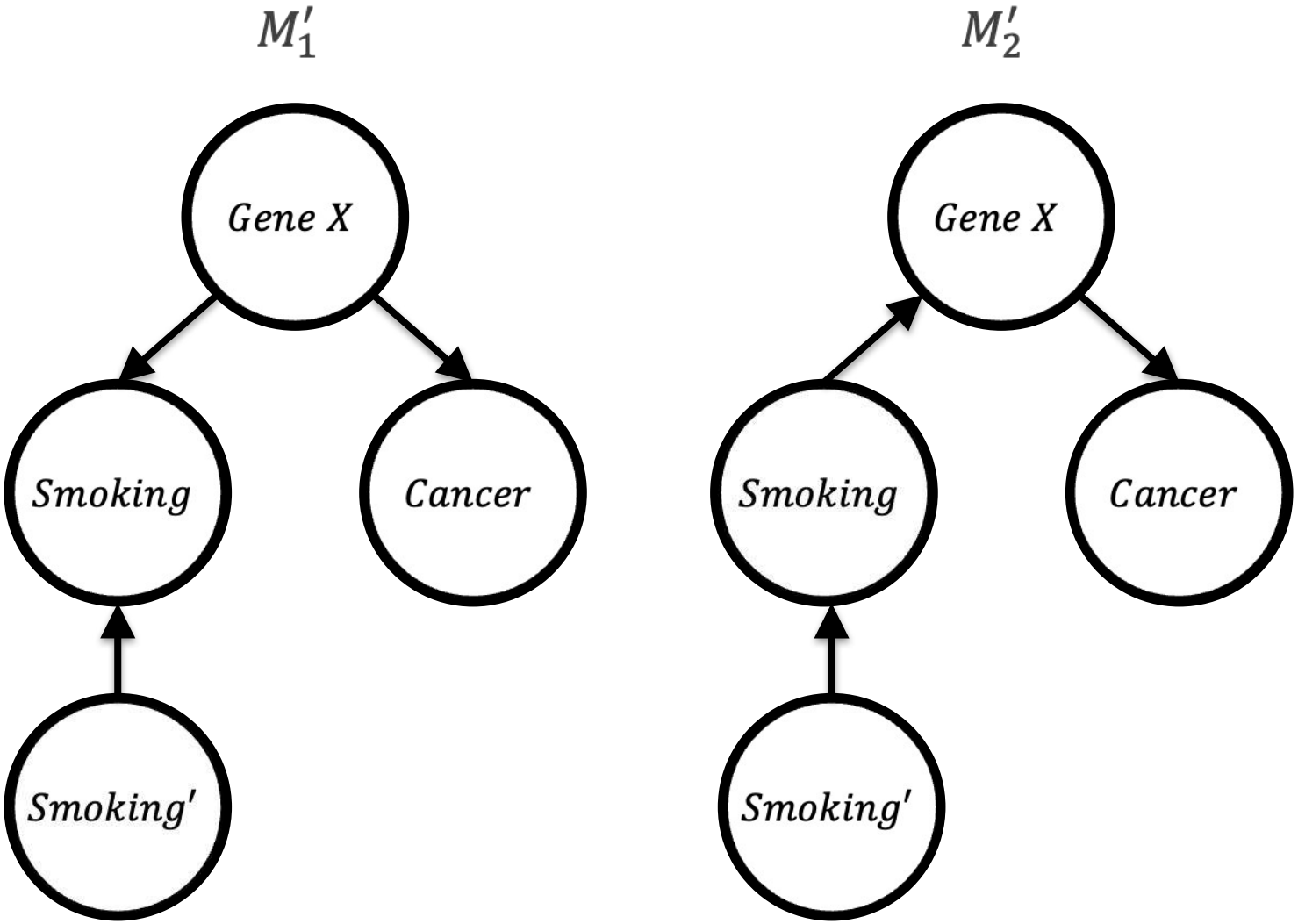
Example of reducibility of $\mathcal{L}_2$ to encriched $\mathcal{L}_1$



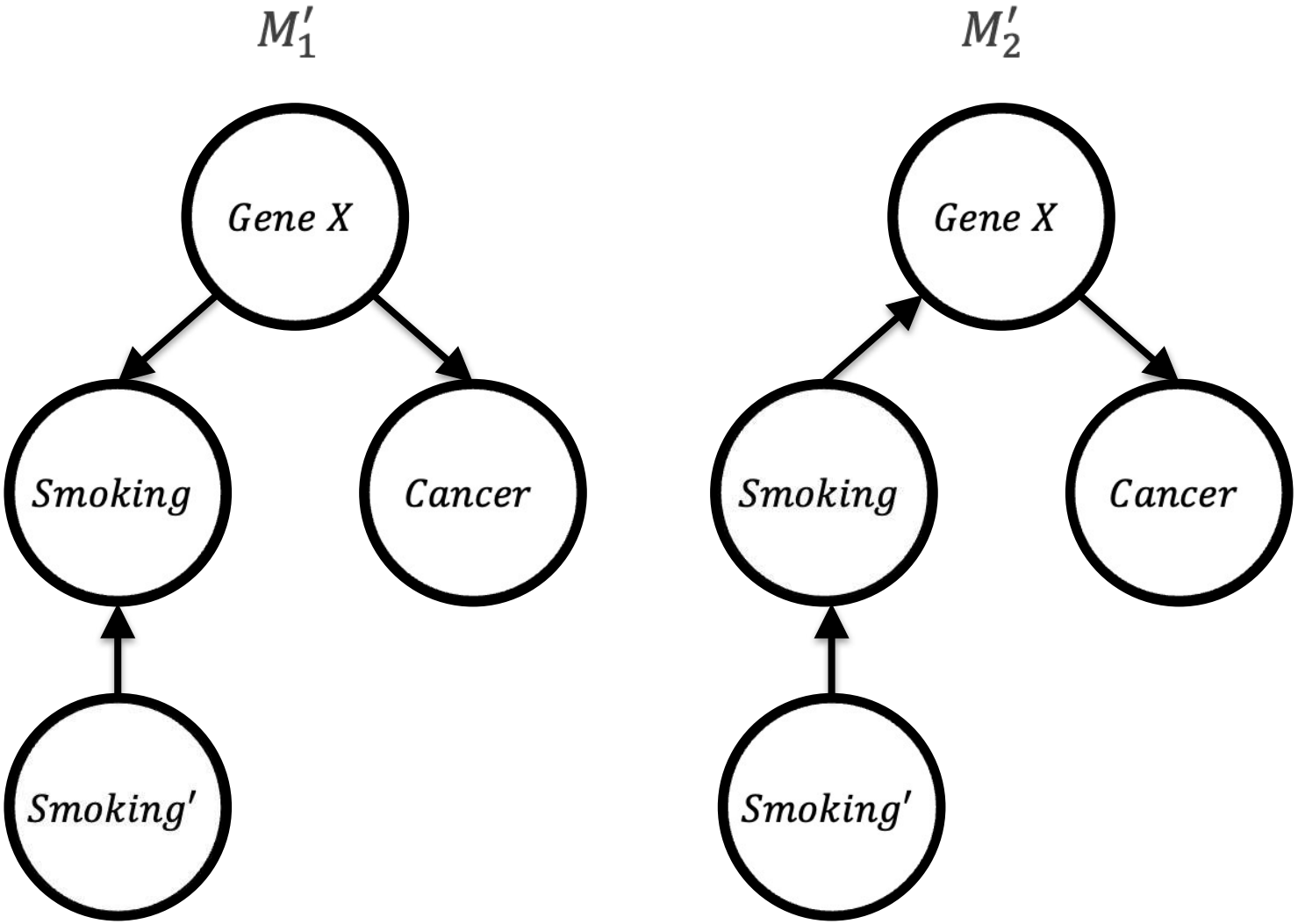
Endogenized interventions on smoking



These graphs are *distinguishable* using facts / observations regarding associations on these variables.



$$P_{M_1}(Cancer|Smoking') \neq P_{M_2}(Cancer|Smoking')$$



This is achieved with conditioning only.

Proof sketch:

1. Preservation of Associational Propositions.

$$P(X|Y=y,M) = P(X|Y=y,Y'=NA,M')$$

2. Equivalence of Interventional Propositions.

$$P(X|do(Y=y),M) = P(X|Y'=y,M')$$

Proof sketch:

1. Preservation of Associational Propositions.

$$P(X|Y = y, M) = P(X|Y = y, Y' = NA, M')$$

2. Equivalence of Interventional Propositions.

$$P(X|do(Y = y), M) = P(X|Y' = y, M')$$

Proof sketch:

1. Preservation of Associational Propositions.

$$P(X|Y = y, M) = P(X|Y = y, Y' = NA, M')$$

2. Equivalence of Interventional Propositions.

$$P(X|do(Y = y), M) = P(X|Y' = y, M')$$

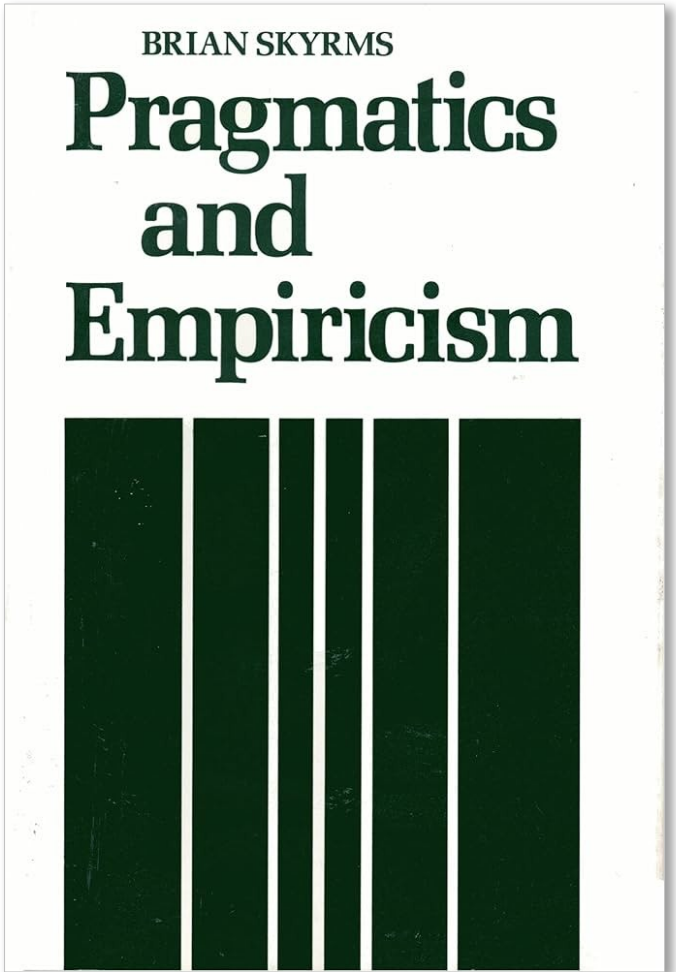
By enriching the agent’s conceptual space, we’ve reduced causal reasoning to probabilistic in the context of SCM.

Let's get subjective...

What are the probabilities in the enriched models?

In SCMs we often treat the probabilities as **objective features** of the world that we're learning about.

But this is disharmonious with a fully subjective Bayesian picture.



Can we reduce the probabilities in SCMs to purely subjective ones?

Let’s step back and consider an analogy.

Chance & de Finetti’s Representation Theorem

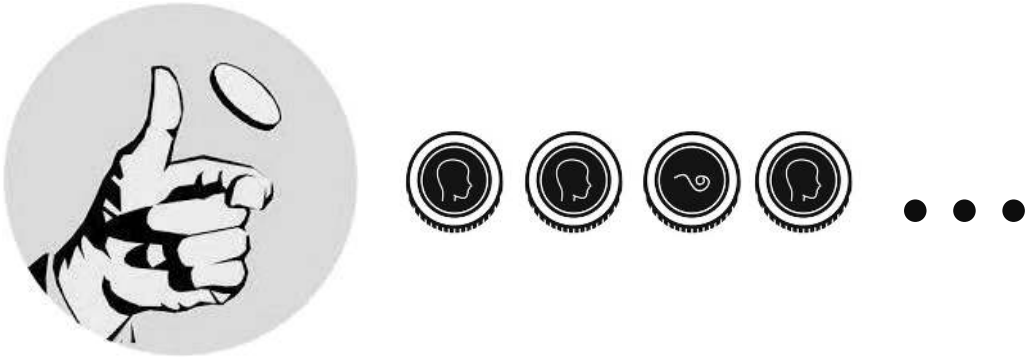
What is chance?

Relative frequencies of sequences of flips





Relative frequencies of sequences of flips



Propensities *in* the coin

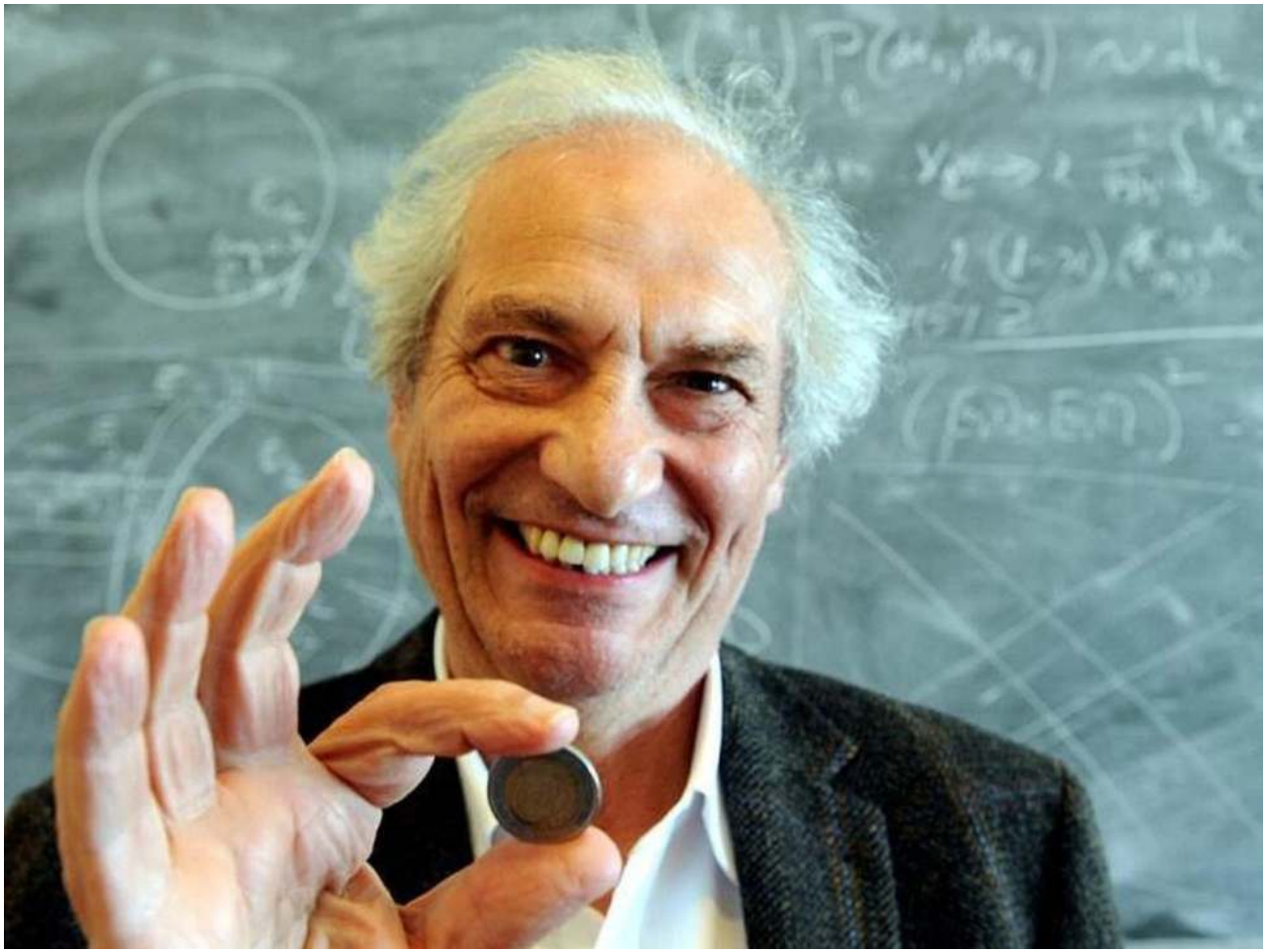
noun

an inclination or natural tendency to behave in a particular way.

Similar

tendency | inclination | predisposition | proneness | proclivity | readiness | susceptibility | liability | disposition | aptness | penchant | leaning | predilection | bent | habit | weakness

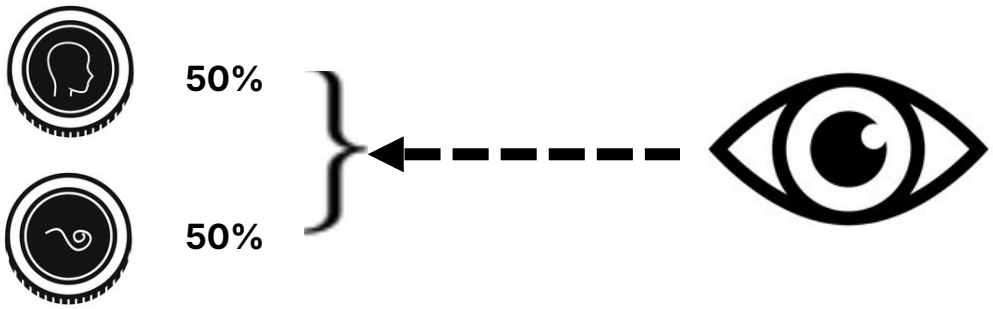
These accounts have problems—among them, that chance appears to **depend on our knowledge**.





Perci flips

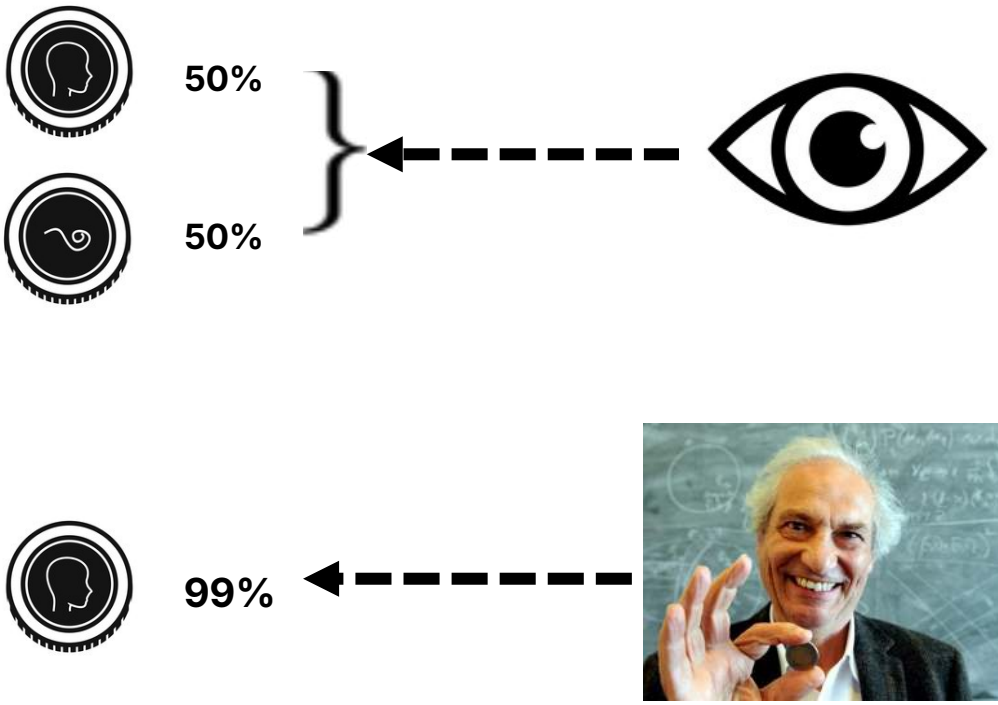


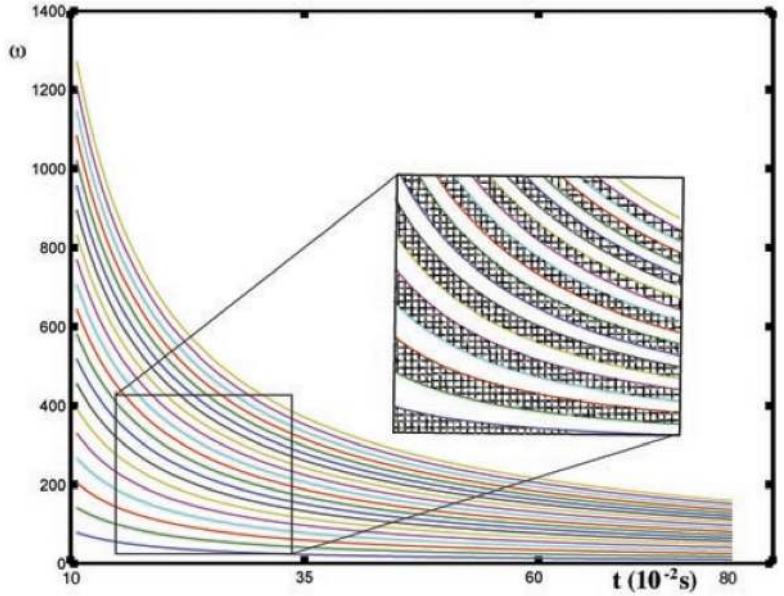


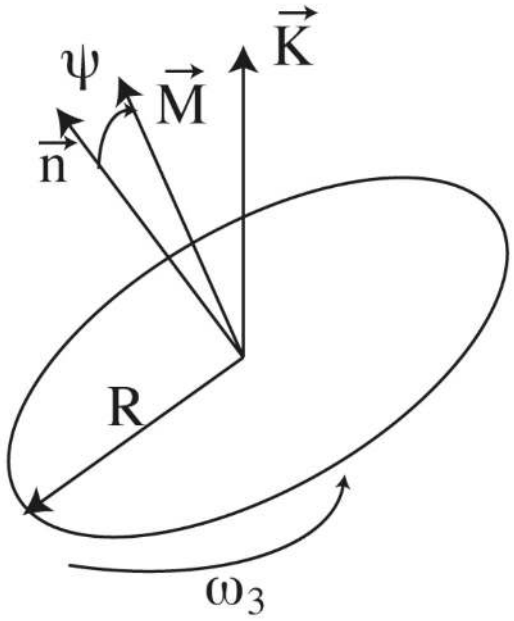
Perci flips



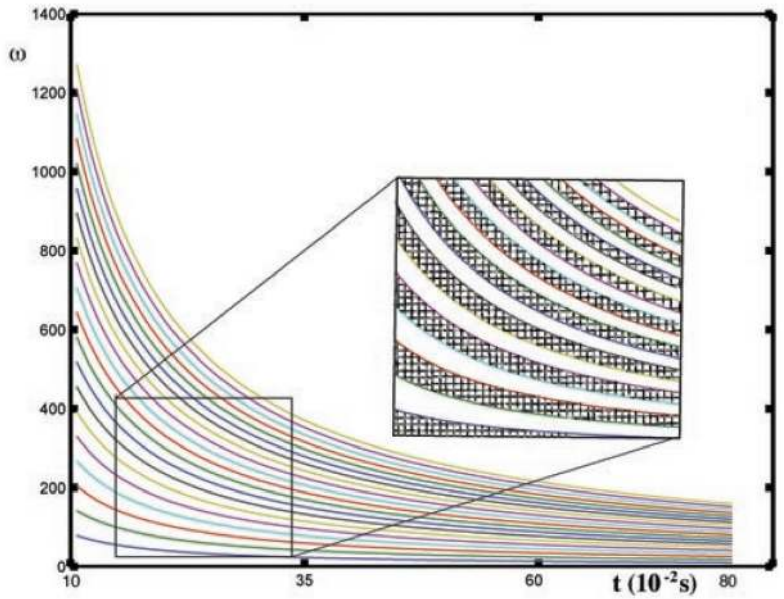
Perci flips

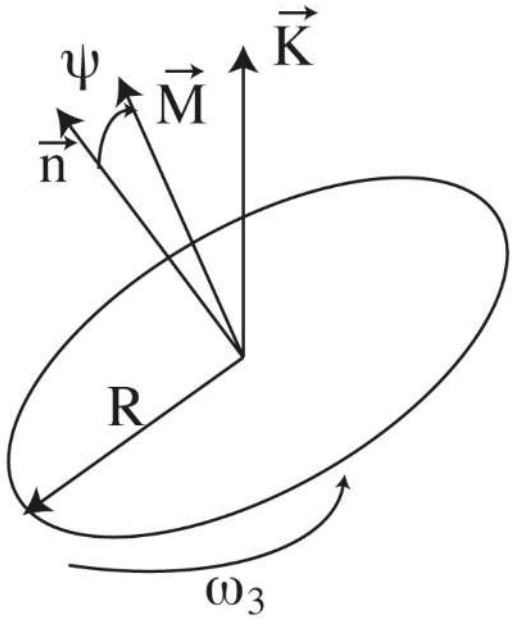






In general, if you knew enough about the physics of a situation, you could calculate the outcome of a coin toss with $P > 0.5$.





Insight. This is true in general.



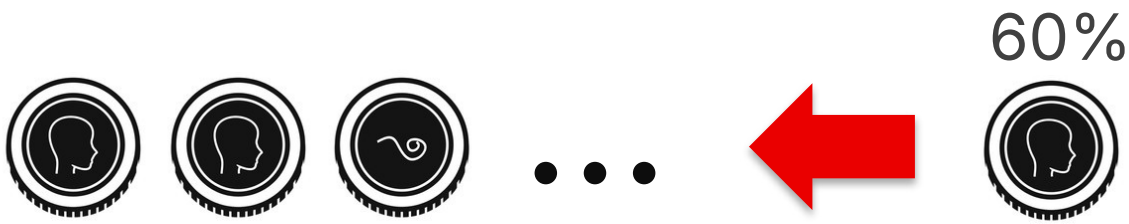
What then is chance θ here?

Famously, de Finneti gave us an account of this...

De Finetti's Representation Theorem

Theorem. Consider an infinite sequence of random variables. If an agent's probabilistic credences P over such sequences is *exchangeable*, then there exists a probability measure μ on the space of probability distributions such that the agent is representable as having a prior over i.i.d. sampling distributions and updating based on observed data.

De Finetti’s Representation Theorem



$$P(X_1, X_2, \ldots, X_n) = \int \prod_{i=1}^{n} P(X_i|\theta)\, d\mu(\theta)$$

Extensions

- Diaconis & Freedman (1980) “**Finite** Exchangeable Sequences”

Extensions

- Diaconis & Freedman (1980) "**Finite** Exchangeable Sequences"
- Diaconis & Freedman (1980) "De Finetti's Theorem for **Markov Chains**"

Extensions

- Diaconis & Freedman (1980) "**Finite** Exchangeable Sequences"
- Diaconis & Freedman (1980) "De Finetti's Theorem for **Markov Chains**"
- Schervish (1995) "A Representation for Sequences of Partially Exchangeable **Arrays**"

Extensions

- Diaconis & Freedman (1980) "**Finite** Exchangeable Sequences"
- Diaconis & Freedman (1980) "De Finetti's Theorem for **Markov Chains**"
- Schervish (1995) "A Representation for Sequences of Partially Exchangeable **Arrays**"
- Freer & Roy (2009) "**Computable** exchangeable sequences have computable de Finetti measures"

Extensions

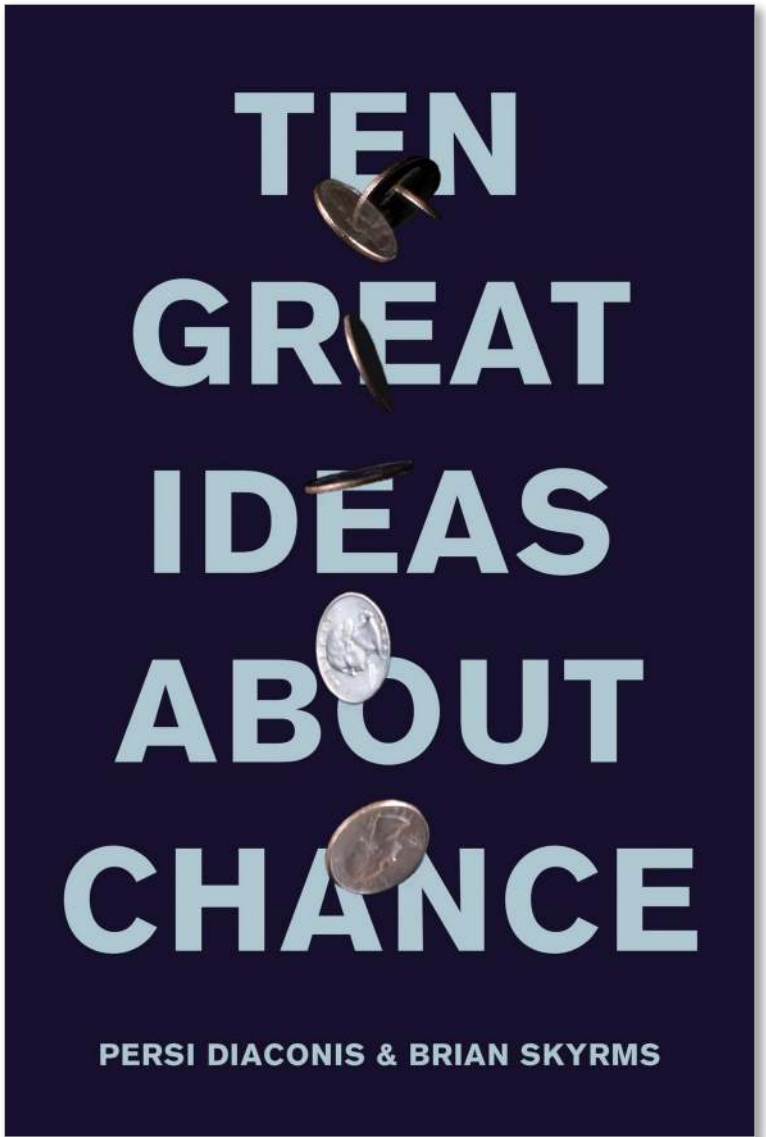
- Diaconis & Freedman (1980) "**Finite** Exchangeable Sequences"
- Diaconis & Freedman (1980) "De Finetti's Theorem for **Markov Chains**"
- Schervish (1995) "A Representation for Sequences of Partially Exchangeable **Arrays**"
- Freer & Roy (2009) "**Computable** exchangeable sequences have computable de Finetti measures"
- Austin & Panchenko (2014) "A **Hierarchical** Version of the de Finetti and Aldous-Hoover Representations"

Extensions

- Diaconis & Freedman (1980) "**Finite** Exchangeable Sequences"
- Diaconis & Freedman (1980) "De Finetti's Theorem for **Markov Chains**"
- Schervish (1995) "A Representation for Sequences of Partially Exchangeable **Arrays**"
- Freer & Roy (2009) "**Computable** exchangeable sequences have computable de Finetti measures"
- Austin & Panchenko (2014) "A **Hierarchical** Version of the de Finetti and Aldous-Hoover Representations"
- Cai, Ackerman, and Freer (2016) "Priors on Exchangeable Directed **Graphs**"

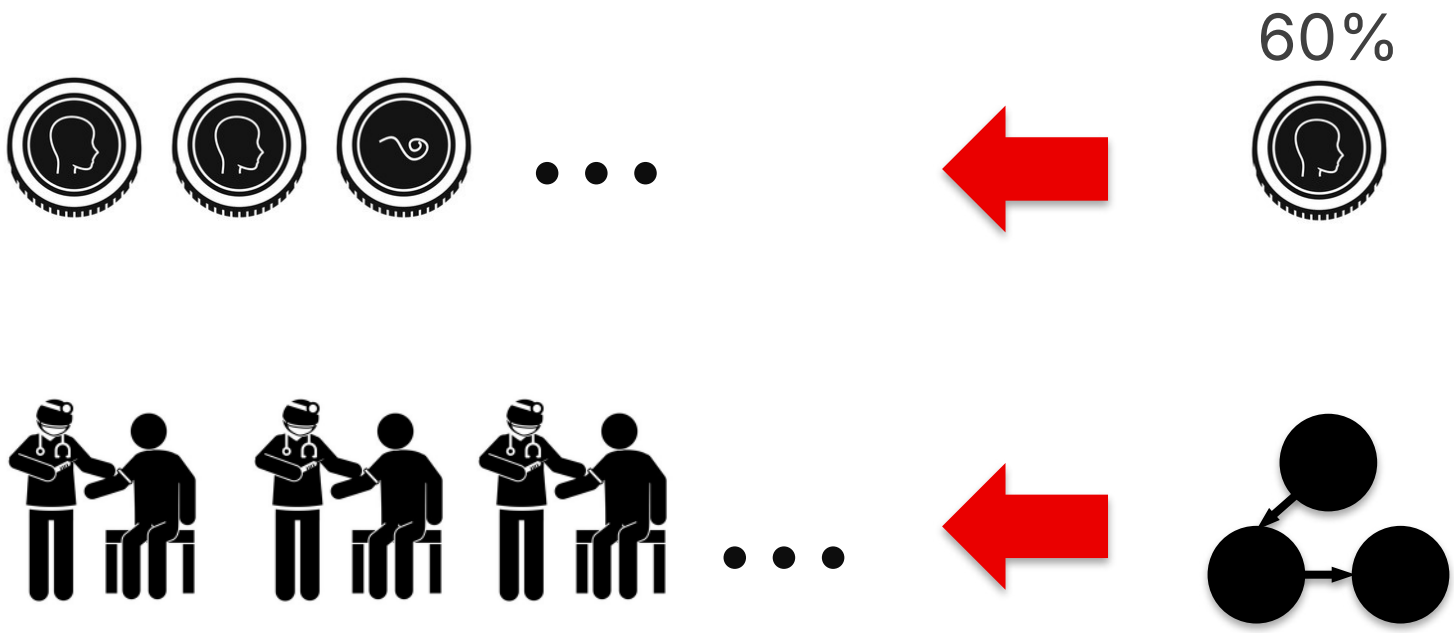
Extensions

- Diaconis & Freedman (1980) "**Finite** Exchangeable Sequences"
- Diaconis & Freedman (1980) "De Finetti's Theorem for **Markov Chains**"
- Schervish (1995) "A Representation for Sequences of Partially Exchangeable **Arrays**"
- Freer & Roy (2009) "**Computable** exchangeable sequences have computable de Finetti measures"
- Austin & Panchenko (2014) "A **Hierarchical** Version of the de Finetti and Aldous-Hoover Representations"
- Cai, Ackerman, and Freer (2016) "Priors on Exchangeable Directed **Graphs**"
- Etc.


TEN
GREAT
IDEAS
ABOUT
CHANCE
PERSI DIACONIS & BRIAN SKYRMS

Our aim here is to do for *causation*
what de Finetti did for *chance*.


60%


60%

A Representation Theorem for Causal Reasoning

Random Matrices

$$\begin{bmatrix} X_1^1 & X_1^2 & \cdots \\ \vdots & \vdots & \\ X_n^1 & X_n^2 & \cdots \end{bmatrix}$$

Random Matrices

$$\begin{bmatrix} X_1^1 & X_1^2 & \cdots \\ \vdots & \vdots & \\ X_n^1 & X_n^2 & \cdots \end{bmatrix}$$

Take the enriched set of variables from our reduction—these include both the original and intervention variables—with repeated observations of those variables.

Random Matrices

$$\begin{bmatrix} X_1^1 & X_1^2 & \cdots \\ \vdots & \vdots & \\ X_n^1 & X_n^2 & \cdots \end{bmatrix}$$

Each column is an observation $D_t = \langle X_1^t, \ldots, X_n^t \rangle$ at time t.

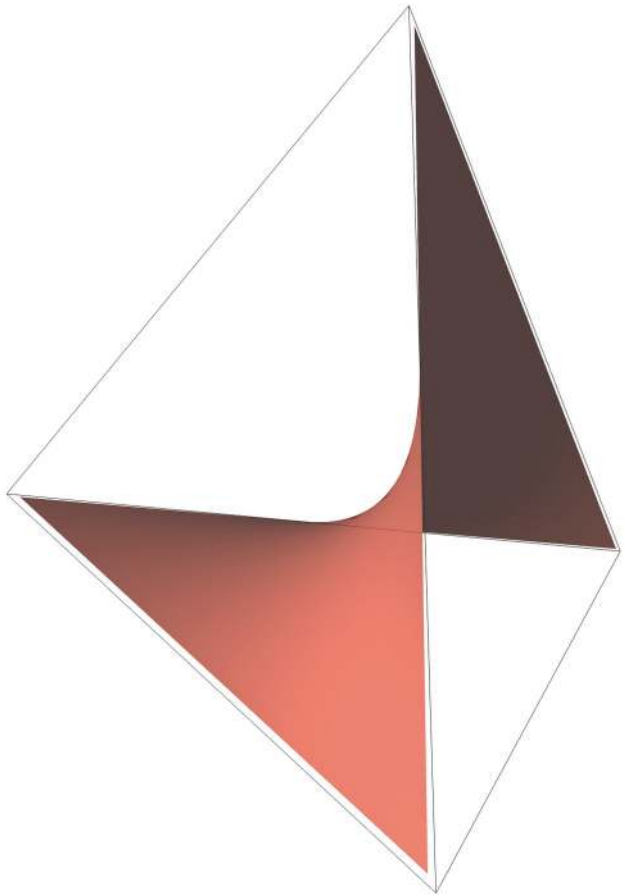
Representation Theorem for Causal Reasoning

Theorem. Let P be an agent's probability function over an infinite sequence of observations $D = \langle D_1, D_2, \ldots \rangle$, where each $D_t = \langle X_1^t, \ldots, X_n^t \rangle$ is a joint observation of all n random variables at time t, and each variable X_i takes values in a finite set. If P satisfies exchangeability over time points, then:

Representation Theorem for Causal Reasoning

1. There exists a probability measure μ over the simplex of joint distributions Δ such that P can be represented as a mixture of i.i.d. sampling distributions.
2. The agent's beliefs and inferences about both causal structure and conditional probability distributions correspond to Bayesian updating on μ given the observed data $D_1, \ldots, D_m$.

Δ



Representation Theorem for Causal Reasoning

1. There exists a probability measure μ over the simplex of joint distributions Δ such that P can be represented as a mixture of i.i.d. sampling distributions.
2. The agent's beliefs and inferences about both "causal structure" and conditional probability distributions correspond to Bayesian updating on μ given the observed data $D_1, \ldots, D_m$.

Proof idea:

Define the product space of the range of the enriched set of variables Ω and the simplex Δ of joint distributions over Ω.

Define an equivalence relation for conditional independence relations inducing a partition $\{R^k\}_{k=1}^K$.

$\{R^k\}_{k=1}^K$ then contains all possible "causal hypotheses" the agent might entertain.

Proof idea:

Define the product space of the range of the enriched set of variables Ω and the simplex Δ of joint distributions over Ω.

Define an equivalence relation for conditional independence relations inducing a partition $\{R^k\}_{k=1}^K$.

$\{R^k\}_{k=1}^K$ then contains all possible "causal hypotheses" the agent might entertain.

Proof idea:

Define the product space of the range of the enriched set of variables Ω and the simplex Δ of joint distributions over Ω.

Define an equivalence relation for conditional independence relations inducing a partition $\{R^k\}_{k=1}^K$.

$\{R^k\}_{k=1}^K$ then contains all possible "causal hypotheses" the agent might entertain.

Let the agent's probability judgments P regarding the observations of the variables D_i be invariant under any permutations π of indices:

$$P(D_1, \dots, D_m) = P\left(D_{\pi(1)}, \dots, D_{\pi(m)}\right).$$

Apply the multinomial version of de Finetti's Theorem to the sequence of observations D to get that exists a μ s.t.

$$P(D_1, \dots, D_m) = \int_{\Delta} \prod_{k=1}^{m} P(D_k) d\mu(\theta).$$

Let the agent's probability judgments P regarding the observations of the variables D_i be invariant under any permutations π of indices:

$$P(D_1, \ldots, D_m) = P\left(D_{\pi(1)}, \ldots, D_{\pi(m)}\right).$$

Apply the multinomial version of de Finetti's Theorem to the sequence of observations D to get that exists a μ s.t.

$$P(D_1, \ldots, D_m) = \int_{\Delta} \prod_{k=1}^{m} P(D_k) d\mu(\theta).$$

$$\int_{\Delta} \prod_{k=1}^{m} P(D_k) d\mu(\theta).$$

μ assigns probability mass to each region R^k, which represents the agent's belief in the corresponding pattern of conditional independencies.

$$\int_{\Delta} \prod_{k=1}^{m} P(D_k) d\mu(\theta).$$

Within each region R^k, the conditional measure $\mu(\theta|R^k)$ represents the agent's uncertainty about the specific parameters given a particular causal structure.

So, when the agent observes data $D_1, \ldots, D_m$, Bayesian updating on μ produces a posterior distribution:

$$\mu(\theta|D_1, \ldots, D_m) \propto \prod_{k=1}^{m} P(D_k|\theta) \times \mu(\theta)$$

This posterior simultaneously updates both:

i. The probability of each region R^k;
(belief about causal structure)

ii. The distribution over parameters within regions.
(belief about specific conditional probabilities)

So, when the agent observes data $D_1, \ldots, D_m$, Bayesian updating on μ produces a posterior distribution:

$$\mu(\theta|D_1, \ldots, D_m) \propto \prod_{k=1}^{m} P(D_k|\theta) \times \mu(\theta)$$

This posterior simultaneously updates both:

i. The probability of each region R^k;
(*belief about causal structure*)

ii. The distribution over parameters within regions.
(*belief about specific conditional probabilities*)

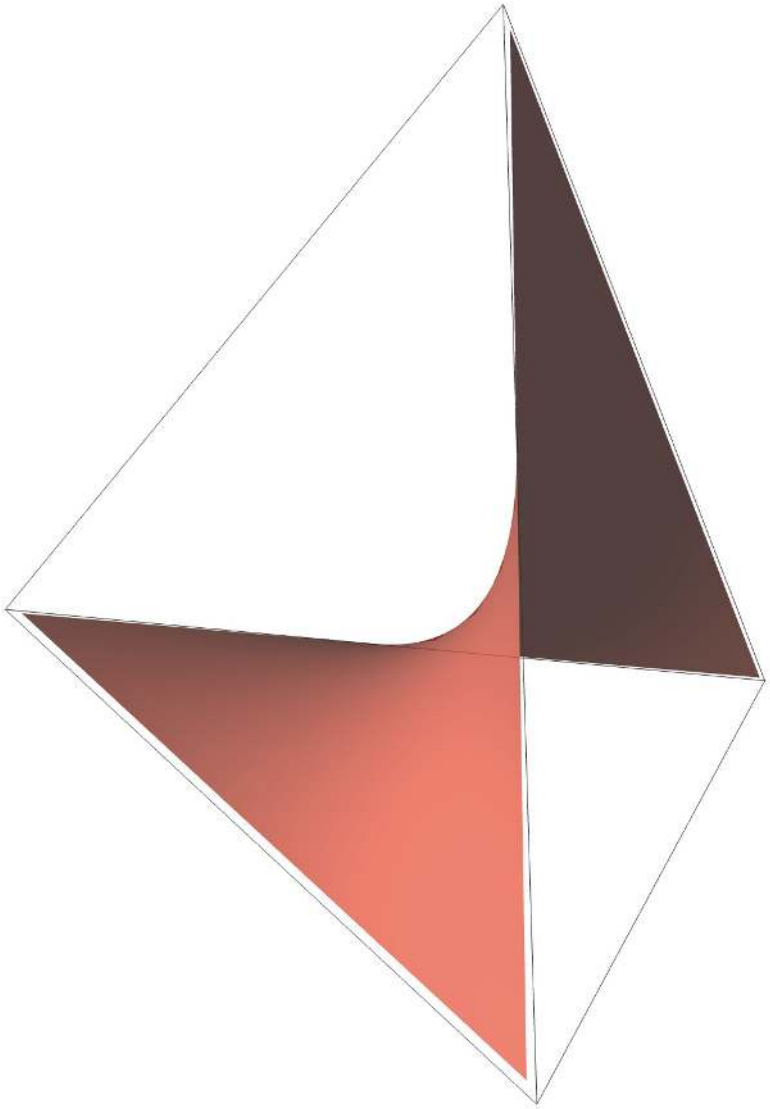
So, when the agent observes data $D_1, \ldots, D_m$, Bayesian updating on μ produces a posterior distribution:

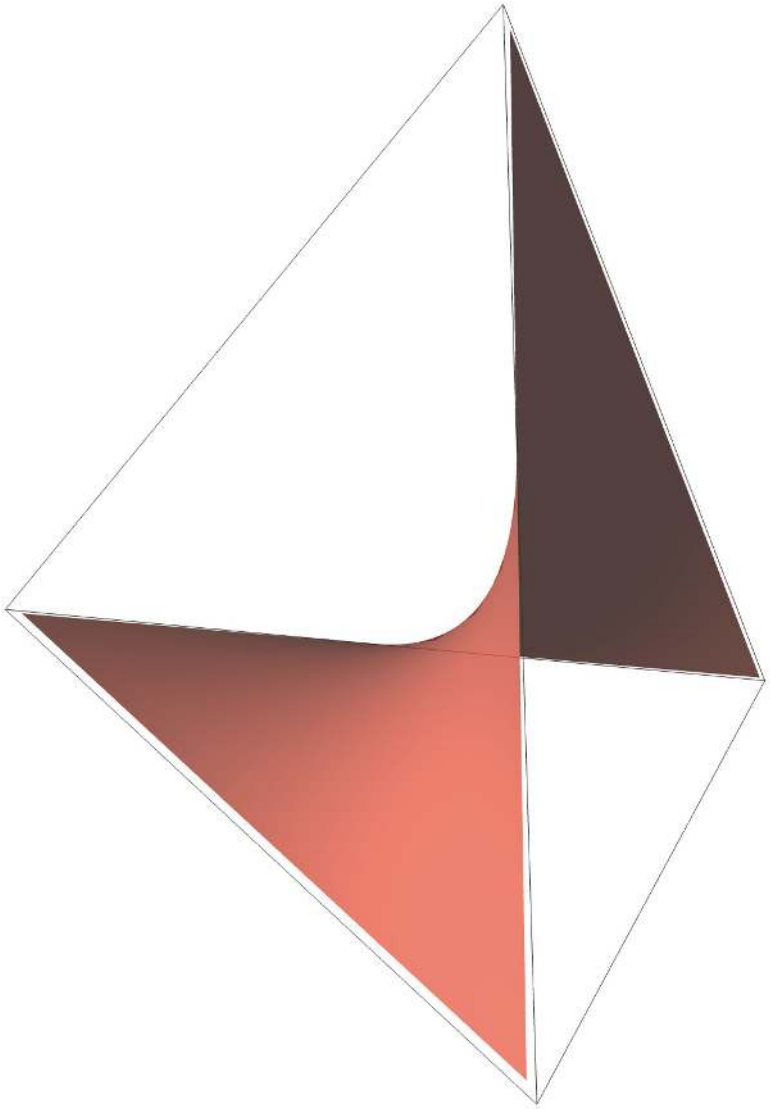
$$\mu(\theta|D_1, \ldots, D_m) \propto \prod_{k=1}^{m} P(D_k|\theta) \times \mu(\theta)$$

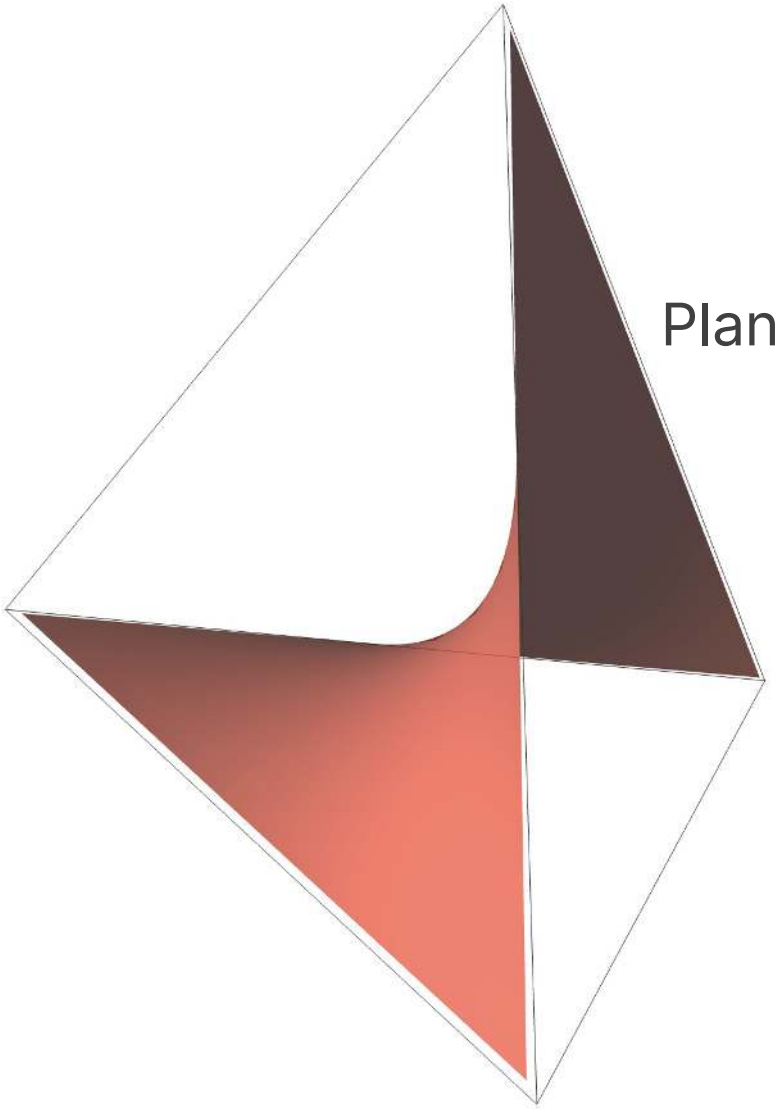
This posterior simultaneously updates both:

i. The probability of each region R^k;
(*belief about causal structure*)

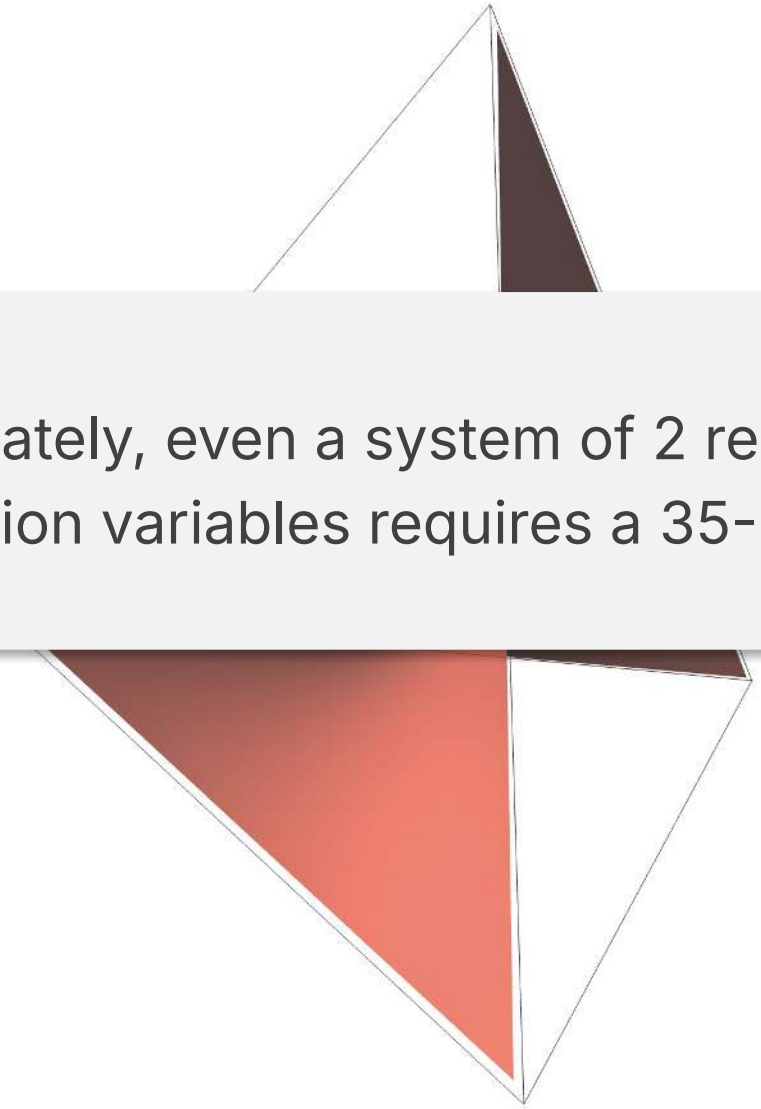
ii. The distribution over parameters within regions.
(*belief about specific conditional probabilities*)



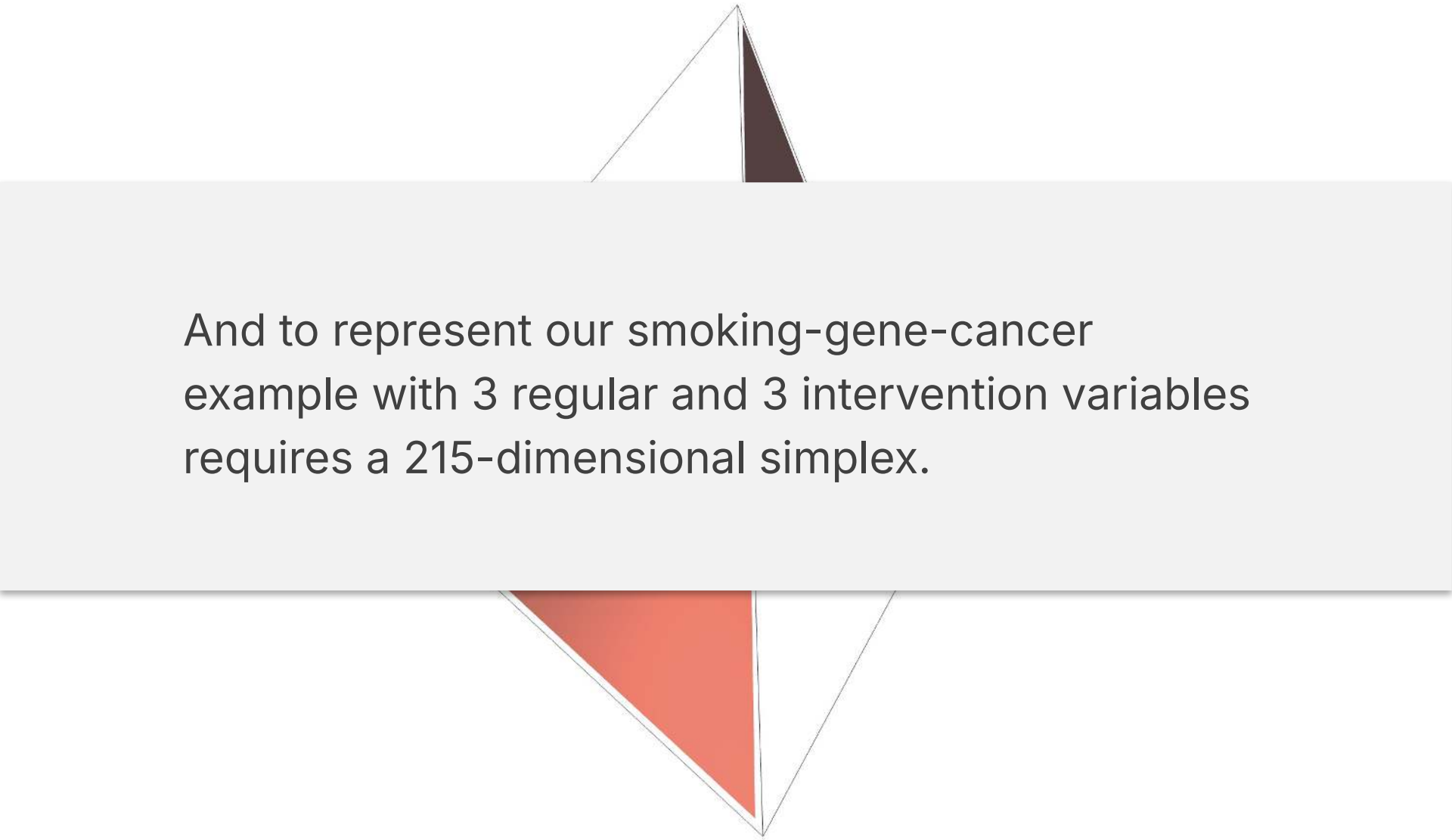




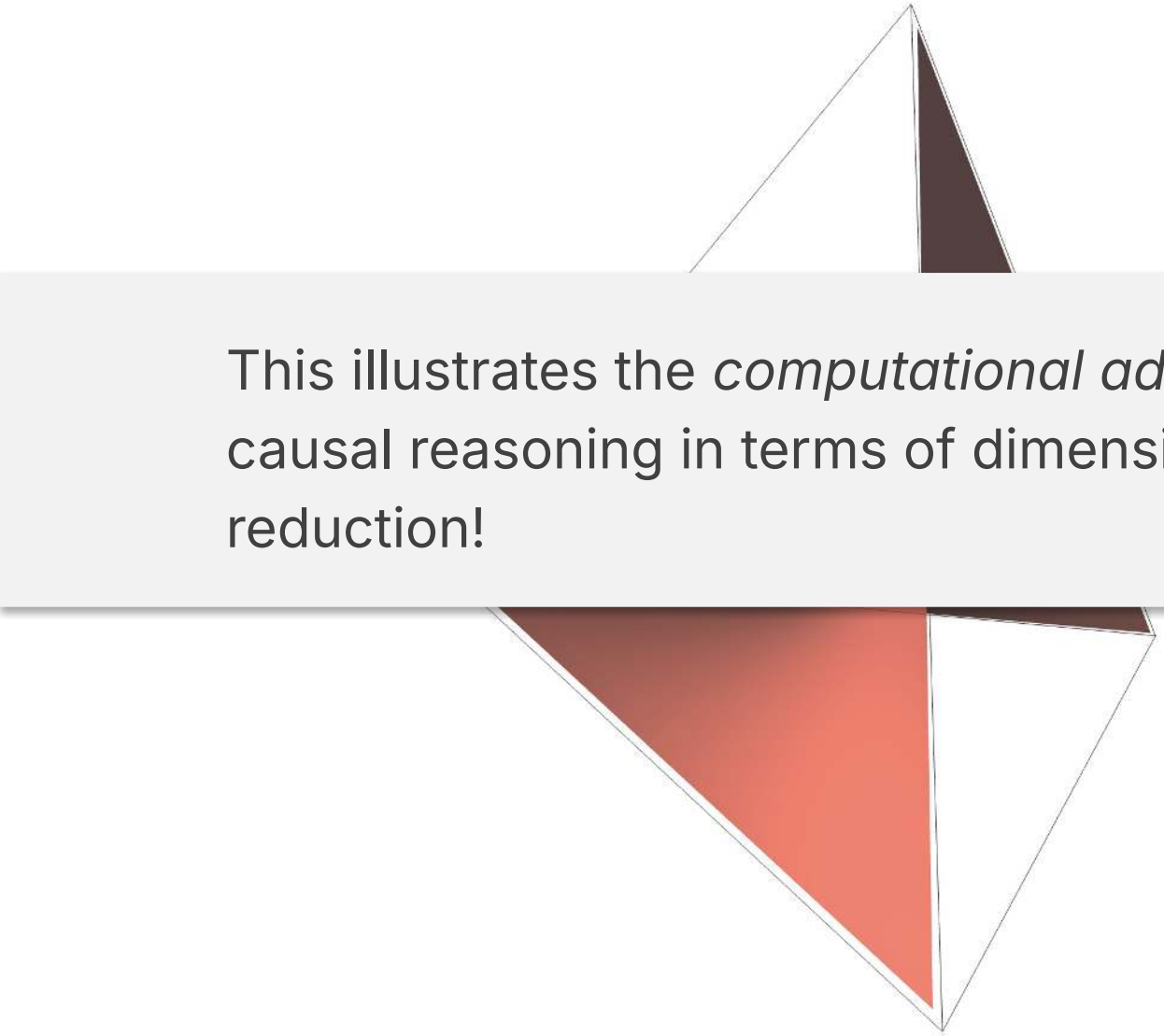
Plane of independence
in a 4-simplex



Unfortunately, even a system of 2 regular and 2 intervention variables requires a 35-D simplex.

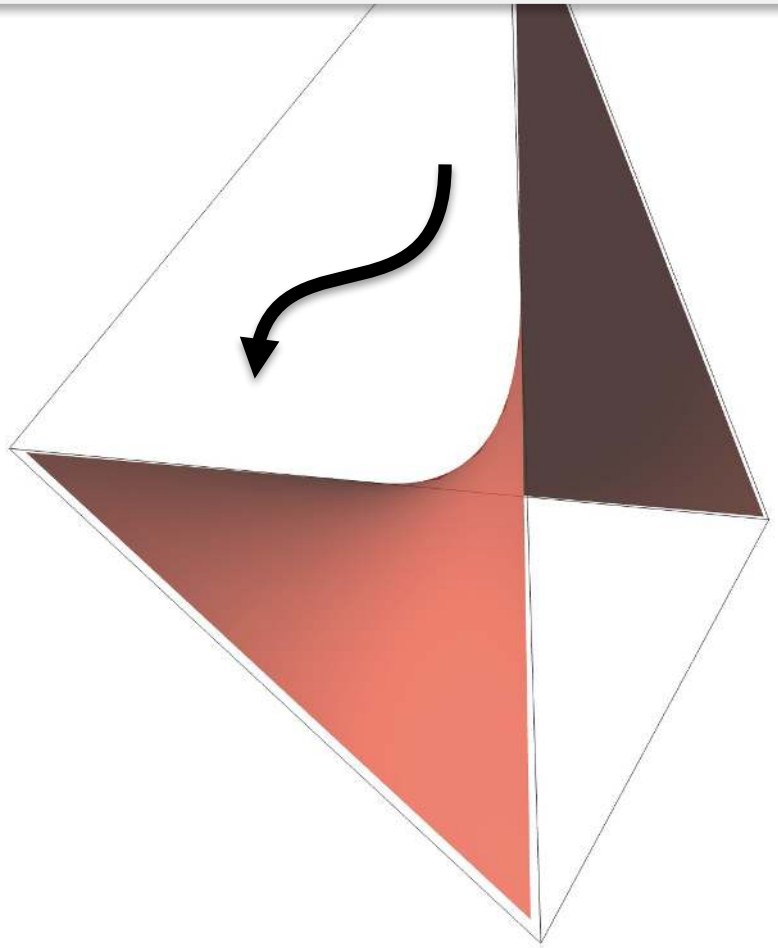


And to represent our smoking-gene-cancer example with 3 regular and 3 intervention variables requires a 215-dimensional simplex.

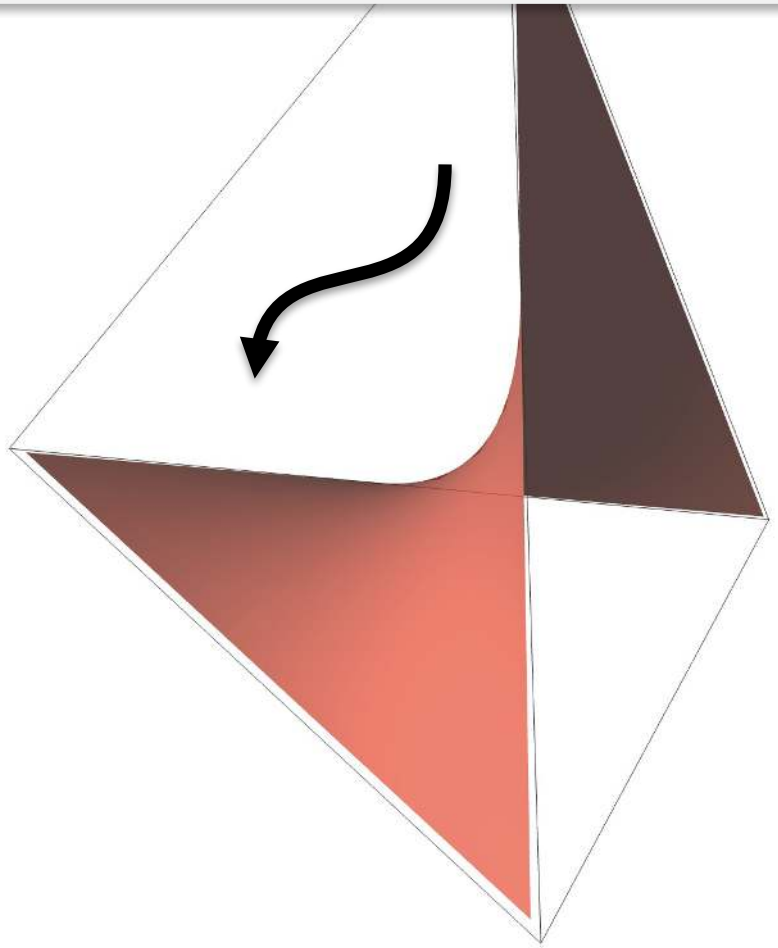


This illustrates the *computational advantage* of causal reasoning in terms of dimensionality reduction!

To learn is...



Causal learning here is a special case of probabilistic learning.



Implications

We have shown that interventional reasoning in the context of SCMs can be *naturalized* in the sense of being reduced to features of our uncertainty.

This doesn’t diminish that causal reasoning is a central pattern of both folk and scientific reasoning.

This also doesn’t decide the status of all types of causation—e.g., causation in fundamental physics, so-called “actual causation,” etc.



The class of continuous timelike curves determines the topology of spacetime*

David B. Malament

Department of Philosophy, University of Chicago, Chicago, Illinois 60637
(Received 23 August 1976)

The title assertion is proven, and two corollaries are established. First, the topology of every past and future distinguishing spacetime is determined by its causal structure. Second, in *every* spacetime the path topology of Hawking, King, and McCarthy codes topological, differential, and conformal structure.

1. SUMMARY

Suppose one has two spacetimes (M,g) and (M',g') together with a bijection $f: M \to M'$, where both f and f^{-1} preserve continuous timelike curves; i.e., if $\gamma: I \to M$ is a continuous timelike curve in (M,g), then $f\circ\gamma: I \to M'$ is a continuous timelike curve in (M',g'); and symmetrically for f^{-1}. We show that f must be a homeomorphism. In this sense the class of continuous timelike curves in spacetime determines its topology.

The result is of interest because, at least in some sense, we directly experience whether events on our worldlines are "close" or not. That experience alone, it appears, allows a complete determination of topological structure. The result also has two consequences which are of independent interest.

It is well known that in all strongly causal spacetimes the Alexandroff topology is equal to the manifold topology.[1] Hence, at least in strongly causal spacetimes, if one knows of all points p and q whether it is possible that a particle travel from p to q, then one can recover the topology of spacetime. The question naturally arises whether the condition of strong causality is necessary for this recovery. We show that it is not. The weaker condition of past and future distinguishability suffices. One has the following result: If (M,g) and (M',g') are past and future distinguishing spacetimes and if $f: M \to M'$ is a causal isomorphism (i.e., a bijection where both f and f^{-1} preserve the causal connectibility relation $\ll$), then f must be a homeomorphism. But we also show that the assertion becomes false if the hypothesis of past *and* future distinguishability is relaxed to that of future distinguishability (or past distinguishability) alone.

A second consequence of our theorem is an improvement of a result of Hawking, King, and McCarthy.[2] They define a *path topology* on spacetimes and prove that, in the presence of strong causality, the path topology "codes" (standard) topological, differential, and conformal structure. We show that their hypothesis of strong causality is unnecessary. Indeed their result is true of *all* spacetimes.

2. STANDARD DEFINITIONS AND RESULTS

sumed to be temporally orientable and endowed with a particular temporal orientation.

Given subsets A and O of M with O open, $I^+(A,O)$ is the set of points q in O such that there exists a future directed smooth timelike curve $\gamma: I \to O$ (where $I \subseteq R$ is connected) and points $t_1, t_2 \in I$ such that $t_1 < t_2$, $\gamma(t_1) \in A$, and $\gamma(t_2) = q$. $I^+(A,O)$ is called the *chronological future of A relative to O*. *The causal future of A relative to O*, $J^+(A,O)$, is the union of $A \cap O$ with the set of points q in O such that there exists a future directed smooth causal curve (i.e., a smooth curve whose tangent vectors are everywhere nonvanishing, nonspacelike, and future directed) $\gamma: I \to O$ and points $t_1, t_2 \in I$ such that $t_1 < t_2$, $\gamma(t_1) \in A$, and $\gamma(t_2) = q$. Finally, the *horismos future of A relative to O*, $E^+(A,O)$, is the set $J^+(A,O) - I^+(A,O)$. These sets have duals $I^-(A,O)$, $J^-(A,O)$, and $E^-(A,O)$ which are defined analogously (substitute past directed curves for future directed curves). $I(A,O)$ is the union $I^+(A,O) \cup I^-(A,O)$. The sets $J(A,O)$ and $E(A,O)$ are defined similarly.

If $A = \{p\}$, we write $I^+(p,O)$ instead of $I^+(A,O)$ and $I^+(p)$ instead of $I^+(p,M)$. Similarly for the other I, J, E sets. The relations $q \in I^+(p,O)$, $q \in J^+(p,O)$, and $q \in E^+(p,O)$ will sometimes be written as $p \ll q(O)$, $p < q(O)$, $p \to q(O)$. Furthermore, $p \ll q(M)$, $p < q(M)$, and $p \to q(M)$ will sometimes be written as $p \ll q$, $p < q$, and $p \to q$.

The I, J, E sets have the following basic properties.[3] If $q \in I^+(p,O)$, then $p \in I^-(q,O)$ and conversely (similarly for the J and E sets). Both $I^+(p,O)$ and $I^-(p,O)$ are open. If $p \ll q(O)$ and $q < r(O)$, then $p \ll r(O)$. Similarly, if $p < q(O)$ and $q \ll r(O)$, then $p \ll r(O)$. If $p \to q(O)$, then, if $\gamma: [0,1] \to 0$ is a future directed smooth causal curve with $\gamma(0) = p$ and $\gamma(1) = q$, γ must be a null geodesic.

An open set O is *convex* iff given any two points p and q in O there is a geodesic $\gamma: [0,1] \to 0$ with $\gamma(0) = p$, $\gamma(1) = q$ and γ is unique (up to reparametrization). If O is an open convex set, then, for all points p in O, $J^+(p,O) = \mathrm{Cl}[I^+(p,O)]$ = the closure in O of $I^+(p,O)$; and $E^+(p,O) = \mathrm{Bnd}[I^+(p,O)]$ = the boundary of $I^+(p,O)$ in O. {These assertions are false in general if O is not convex. But $J^+(p,O) \subseteq \mathrm{Cl}[I^+(p,O)]$ and $E^+(p,O) \subseteq \mathrm{Bnd}[I^+(p,O)]$ are always true.} Dual assertions hold for J^- and E^-. The open convex sets form a basis for the manifold

0020-7748/78/0615-0293 $02.00/0



Synthese
DOI 10.1007/s11229-009-9497-9

Actual causation: a stone soup essay

Clark Glymour · David Danks · Bruce Glymour · Frederick Eberhardt · Joseph Ramsey · Richard Scheines · Peter Spirtes · Choh Man Teng · Jiji Zhang

Received: 31 March 2008 / Accepted: 1 March 2009
© Springer Science+Business Media B.V. 2009

Abstract We argue that current discussions of criteria for actual causation are ill-posed in several respects. (1) The methodology of current discussions is by induction from intuitions about an infinitesimal fraction of the possible examples and counterexamples; (2) cases with larger numbers of causes generate novel puzzles; (3) “neuron” and causal Bayes net diagrams are, as deployed in discussions of actual causation, almost always ambiguous; (4) actual causation is (intuitively) relative to an initial system state since state changes are relevant, but most current accounts ignore state changes through time; (5) more generally, there is no reason to think that philosophical judgements about these sorts of cases are normative; but (6) there is a dearth of relevant psychological research that bears on whether various philosophical accounts are descriptive. Our skepticism is not directed towards the possibility of a

C. Glymour (✉) · D. Danks
Carnegie Mellon University and Florida Institute for Human and Machine Cognition, Pittsburgh, PA, USA
e-mail: cg09@andrew.cmu.edu

B. Glymour
Kansas State University, Manhattan, KS, USA

Some have claimed that because of the claimed irreducibility, purely associational LLMs cannot—*in principle*—reason causally.

“Large language models and other models employing associational reasoning cannot reason causally.”

—Zečević et al. (2023) “Causal Parrots: Large Language Models May Talk Causality But Are Not Causal” *Transactions in Machine Learning Research*

Our results provide evidence against such claims.

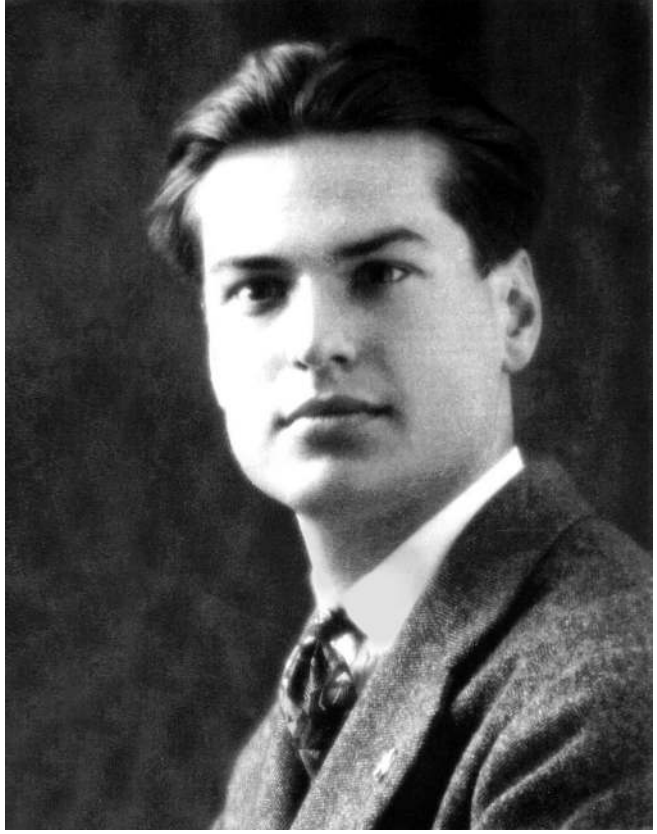
We provide further evidence that—as many thinkers have hypothesized (Hume, 1739; Russell 1913; Reichenbach, 1956; Suppes, 1970)—that causal reasoning performs a different function, perhaps as a heuristic with computationally and pragmatically desirable properties—e.g., dimensionality reduction—for boundedly rational agents.

If our picture is right, **causal reasoning** in a smaller algebra **helps us approximate fully Bayesian reasoning** in a larger algebra.

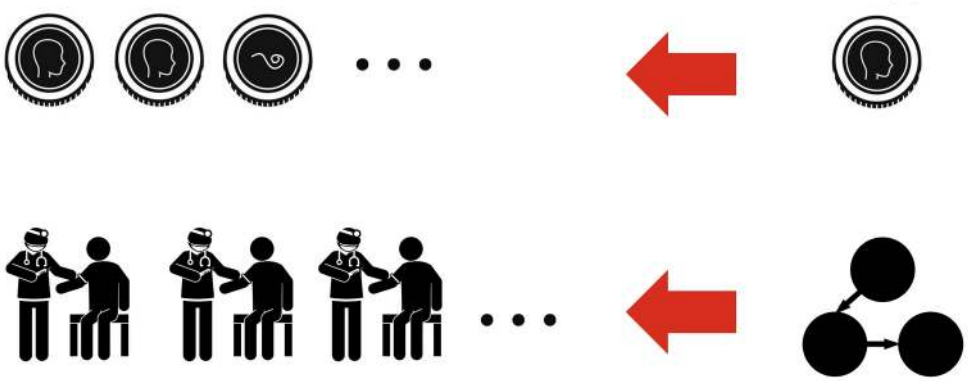
If our picture is right, **causal reasoning** in a smaller algebra **helps us approximate fully Bayesian reasoning** in a larger algebra.

This suggest that there may be *other patterns* of effective approximation.

This can also help explain our causal intuitions in decision-making contexts, while actually **supporting evidential decision theory**.



Our findings advance the naturalizing project articulated by de Finetti (1931), Diaconis and Freedman (1980), and Skyrms (1984), among others.



Just as their results show us how probabilistic judgments can account for chances without reifying them as external entities, we show how probabilistic judgments can also account for causal relations *without reifying them as external entities*

Whatever else causation might be, in the context of SCMs, *it is indeed reducible*.





Happy Halloween.

Carnegie Mellon University

INSTITUTE FOR COMPLEX SOCIAL DYNAMICS