A Dynamic Interpretation of Structural Causal Models

Dean McHugh

Institute of Logic, Language and Computation
and Department of Philosophy,
University of Amsterdam

COCOA Seminar
Friday, 23 February 2024

Institute for Logic, Language and Computation

University of Amsterdam

Plan

1. Structural causal models
2. Motivations
   - Model relativity
   - Representing dependence between properties
   - Dense causal chains
   - Evaluation time
3. A new semantics of conditionals
4. A dynamic interpretation of structural causal models
   - Deciding which structural causal models are correct
   - Further examples of dynamic interpretations
5. Previous hints at the dynamic interpretation
6. Interventions as a special case of sufficiency

Interpreting a formalism

- Probability theory

Interpreting a formalism

- Probability theory
  - **Formalism:** Kolmogorov's axioms (Kolmogorov 1933)

Interpreting a formalism

- Probability theory
  - **Formalism:** Kolmogorov's axioms (Kolmogorov 1933)
  - **Interpretation:** Bayesian, frequentist, propensity, ... (Hájek 2023)

Interpreting a formalism

- Probability theory
  - **Formalism:** Kolmogorov's axioms (Kolmogorov 1933)
  - **Interpretation:** Bayesian, frequentist, propensity, ... (Hájek 2023)
- Quantum mechanics

Interpreting a formalism

- Probability theory
  - **Formalism:** Kolmogorov's axioms (Kolmogorov 1933)
  - **Interpretation:** Bayesian, frequentist, propensity, ... (Hájek 2023)
- Quantum mechanics
  - **Formalism:** Hilbert spaces, the Schrödinger equation, ...

Interpreting a formalism

- Probability theory
  - **Formalism:** Kolmogorov's axioms (Kolmogorov 1933)
  - **Interpretation:** Bayesian, frequentist, propensity, ... (Hájek 2023)
- Quantum mechanics
  - **Formalism:** Hilbert spaces, the Schrödinger equation, ...
  - **Interpretation:** Copenhagen, many worlds, ...

Interpreting a formalism

- Probability theory
  - **Formalism:** Kolmogorov's axioms (Kolmogorov 1933)
  - **Interpretation:** Bayesian, frequentist, propensity, ... (Hájek 2023)
- Quantum mechanics
  - **Formalism:** Hilbert spaces, the Schrödinger equation, ...
  - **Interpretation:** Copenhagen, many worlds, ...
- Structural causal models

Interpreting a formalism

- Probability theory
  - **Formalism:** Kolmogorov's axioms (Kolmogorov 1933)
  - **Interpretation:** Bayesian, frequentist, propensity, ... (Hájek 2023)
- Quantum mechanics
  - **Formalism:** Hilbert spaces, the Schrödinger equation, ...
  - **Interpretation:** Copenhagen, many worlds, ...
- Structural causal models
  - **Formalism:** Variables, structural equations

Interpreting a formalism

- Probability theory
  - **Formalism:** Kolmogorov's axioms (Kolmogorov 1933)
  - **Interpretation:** Bayesian, frequentist, propensity, ... (Hájek 2023)
- Quantum mechanics
  - **Formalism:** Hilbert spaces, the Schrödinger equation, ...
  - **Interpretation:** Copenhagen, many worlds, ...
- Structural causal models
  - **Formalism:** Variables, structural equations
  - **Interpretation:** Mechanisms (Pearl 2000, Halpern and Pearl 2005), dynamic interpretation

Plan

1. Structural causal models
2. Motivations
   - Model relativity
   - Representing dependence between properties
   - Dense causal chains
   - Evaluation time
3. A new semantics of conditionals
4. A dynamic interpretation of structural causal models
   - Deciding which structural causal models are correct
   - Further examples of dynamic interpretations
5. Previous hints at the dynamic interpretation
6. Interventions as a special case of sufficiency

Structural causal models

Definition

A *structural causal model* is tuple

$$(U, V, R, F)$$

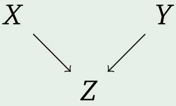
where

- U and V are disjoint sets of variables, called *exogenous* and *endogenous*, respectively.
- R assigns to each variable in $U \cup V$ a set of values.
- F assigns to each endogenous variable $X \in V$ a function $f_X : R(PA_X) \to R(X)$ where $PA_X \subseteq U \cup V \setminus \{X\}$.

(Pearl 2009, p. 203)

Example

$Z = X \vee Y$



$V = \{X, Y, Z\}$
$R(U) = \{0, 1\}$ for all $U \in V$
$F_X(Y, Z) = 1$ iff $Y = 1$ or $Z = 1$

Figure: Structural causal model of an OR-gate.

Interventions in structural causal models

Let $M = (V, E, F)$ be a structural causal model

Definition (Interventions as model surgery)

$M_{X=x}$ is the model $(V, E, F_{X=x})$ which results from replacing the equation for X in M with $X = x$ (that is, $F_{X=x} := (F \setminus \{F_X\}) \cup \{F'_X\}$ where $F'_X(y_1, y_2, \dots) = x$ for any values $y_1, y_2, \dots$ of X's parents).

Interventions in structural causal models

Let $M = (V, E, F)$ be a structural causal model

Definition (Interventions as model surgery)

$M_{X=x}$ is the model $(V, E, F_{X=x})$ which results from replacing the equation for X in M with $X = x$ (that is, $F_{X=x} := (F \setminus \{F_X\}) \cup \{F'_X\}$ where $F'_X(y_1, y_2, \dots) = x$ for any values $y_1, y_2, \dots$ of X's parents).

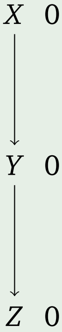
Definition (Truth conditions for interventions)

Let M be a structural causal model and u a setting of the exogenous variables.

$$M, u \models [X \leftarrow x]Y = y \quad \text{iff} \quad M_{X=x}, u \models Y = y$$

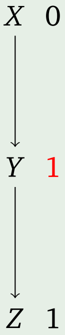
Example of an intervention: a chain

Example



$$Y = X$$
$$Z = Y$$

Example



$$Y = 1$$
$$Z = Y$$

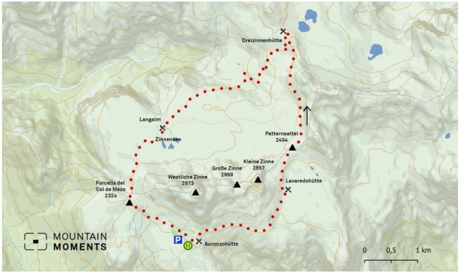
Let M be the model above and $u = (0, 0, 0)$.

$$M, u \models [Y = 1]X = 0$$
$$M, u \models [Y = 1]Z = 1$$

Plan

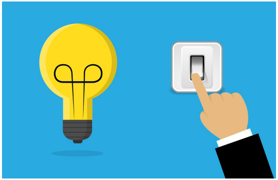
Plan

1. Structural causal models
2. Motivations
   - Model relativity
   - Representing dependence between properties
   - Dense causal chains
   - Evaluation time
3. A new semantics of conditionals
4. A dynamic interpretation of structural causal models
   - Deciding which structural causal models are correct
   - Further examples of dynamic interpretations
5. Previous hints at the dynamic interpretation
6. Interventions as a special case of sufficiency



The map

and the territory



$$S \to L \qquad L = S$$

(a) ✓

$$L \to S \qquad S = L$$

(b) ✗

Figure: How do we decide that model (a) is correct and model (b) is incorrect?

An engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the right-hand track, instead of the left. Since the tracks reconverge up ahead, the train arrives at its destination all the same.

(Hall 2000, p. 205)

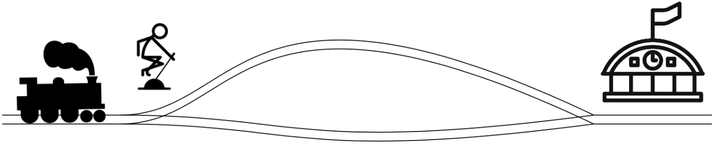


Figure: Hall's switching scenario.

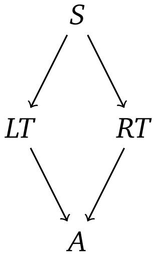
(1) The engineer flipping the switch is a cause of the train reaching the station.

Two models of the switching scenario

$$T = \begin{cases} L & \text{if } S \\ R & \text{if } \neg S \end{cases}$$

$$A = \begin{cases} 1 & \text{if } T = L \vee T = R \\ 0 & \text{otherwise} \end{cases}$$

(a) One-variable model



$$LT = S$$
$$RT = \neg S$$
$$A = LT \vee RT$$

(b) Two-variable model

Halpern (2016) presents a semantics of *is a cause of*. On his semantics,

(1) The engineer flipping the switch is a cause of the train reaching the station.

- is false in the one-variable model
- but true in the two-variable model

(1) The engineer flipping the switch is a cause of the train reaching the station.

- Halpern's semantics of *is a cause of* is highly sensitive to the choice of model.
- The correctly predict that (1) is false, Halpern must rule out the two-variable model. How to do so?

Manipulation

"No Causation without Manipulation" (Holland 1986).

> *one says "A causes B" in cases where one could produce an event or state of the A sort as a means to producing one of the B sort.*
>
> *(Gasking 1955, p. 485)*

> *The paradigmatic assertion in causal relationships is that manipulation of a cause will result in the manipulation of an effect. . . . Causation implies that by varying one factor I can make another vary.*
>
> *(Cook and Campbell 1979, p. 36)*

> *. . . an event A is a cause of a distinct event B just in case bringing about the occurrence of A would be an effective means by which a free agent could bring about the occurrence of B.*
>
> *(Menzies and Price 1993, p. 187)*

Challenges to the manipulationist view

Well-known challenges to this view (Woodward 2016).

- Particularly suited to causation at the human scale, but does not generalise well beyond that (see e.g. Pearl 2000, p. 361).

Challenges to the manipulationist view

Well-known challenges to this view (Woodward 2016).

- Particularly suited to causation at the human scale, but does not generalise well beyond that (see e.g. Pearl 2000, p. 361).
- Consider: The big bang caused stars to form.

Challenges to the manipulationist view

Well-known challenges to this view (Woodward 2016).

- Particularly suited to causation at the human scale, but does not generalise well beyond that (see e.g. Pearl 2000, p. 361).
- Consider: The big bang caused stars to form.
- Perhaps we extend our concept of agency via imagination: If we *had* an effective means to bring about the big bang, ...

Challenges to the manipulationist view

Well-known challenges to this view (Woodward 2016).

- Particularly suited to causation at the human scale, but does not generalise well beyond that (see e.g. Pearl 2000, p. 361).
- Consider: The big bang caused stars to form.
- Perhaps we extend our concept of agency via imagination: If we *had* an effective means to bring about the big bang, ...
- Remaining question: What features of the scenario do we use to decide what would happen if we had the effective means to bring about the big bang?

> *Each parent–child relationship in the network represents a stable and autonomous physical mechanism.*
>
> *(Pearl 2000, p. 22)*

> *Each equation represents a distinct mechanism (or law) in the world, one that may be modified (by external actions) without altering the others.*
>
> *(Halpern and Pearl 2005, p. 847)*

Shifts the question:

- from *When is a causal model correct?*
- to *When does a causal model correctly represent the mechanisms?*

Some sequences of states are lawful (or, nomically possible) and others are not:

Figure: A lawful sequence.

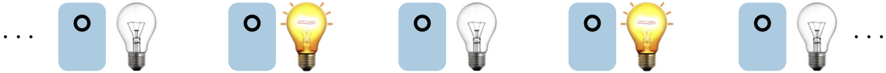


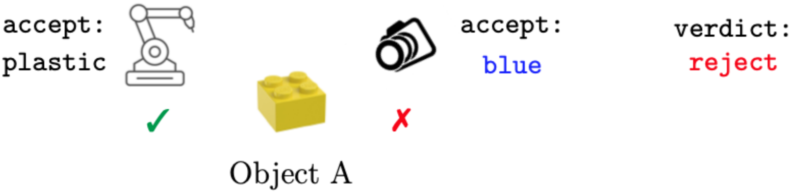
Figure: An unlawful sequence.

Proposal

1. A structural causal model states classifies some sequences of states as lawful and others as unlawful.
2. The dynamic interpretation will tell us, for any model, which sequences of states it classifies as lawful and which as unlawful.
3. A structural causal model is correct just in case
   - the sequences it classifies as lawful are lawful, and
   - the sequences it classifies as unlawful are unlawful.

Plan

- We are in a recycling plant, sorting objects based on material and colour.
- We set the material checker to accept plastic and the camera to accept blue objects.
- An item is accepted just in case it passes both the material and colour check.

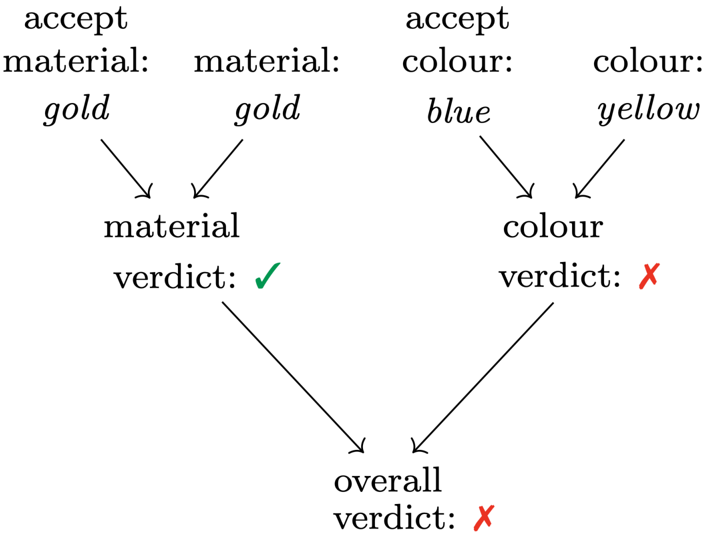


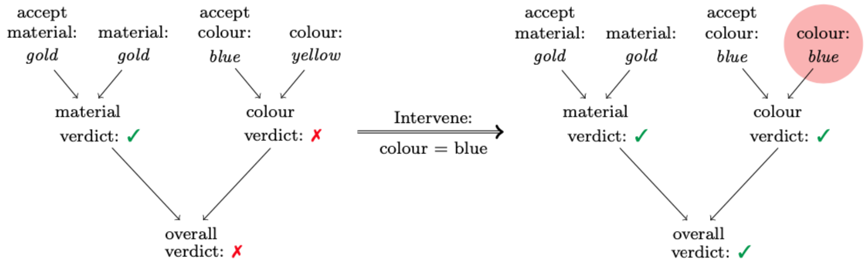
(2) If object A had been blue, it would have been accepted.

- Now we set the material checker to accept pure gold.
- As before,the camera is set to accept blue objects.

Object B

(3) If object B had been blue, it would have been accepted.


accept material: *gold*
material: *gold*
accept colour: *blue*
colour: *yellow*
material verdict: ✓
colour verdict: ✗
overall verdict: ✗



(4) If object B had been blue, it would have been accepted.

Plan

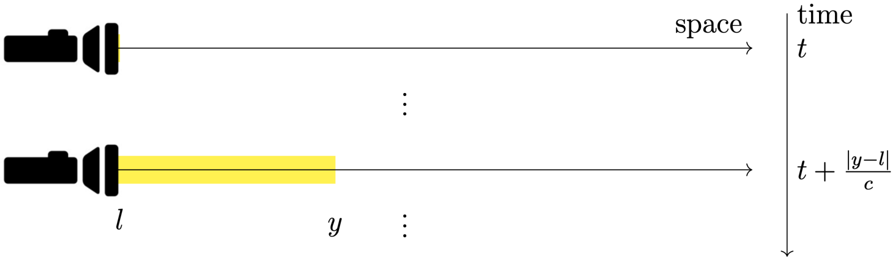


Figure: A dense causal chain.

Structural causal models cannot represent dense causal chains (see McHugh 2023, Proposition 6.7.2).

Definition (Dense dependence)

Let M be a structural causal model and Y a variable of M. We say dependence is *dense* at Y iff for every parent X of Y there is a parent Z of Y such that

$$f_Y(\ldots, x, z, \ldots) = f_Y(\ldots, x', z, \ldots)$$

for all values x, x' of X and value z of Z.

Definition

We say Y *depends on* X in M just in case there is a setting of Y's parents such that changing the value of X results in a change in the value of Y: $f_Y(\ldots, x, \ldots) \neq f_Y(\ldots, x', \ldots)$ for some values x, x' of X.

Proposition

No structural causal model has a variable Y such that

1. *Y depends on some variables*
2. *Dependence is dense at Y.*

Proof.

Suppose such a structural causal model existed. By (1), there is variable X with values x, x' and values o of the parents of Y other than X such that $f_Y(x, o) \neq f_Y(x', o)$. And by (2), there is a parent Z of Y such that $f_Y(x, z, o_{-z}) = f_Y(x', z, o_{-z})$, where o_{-z} are the values in o other than z. A contradiction follows:

$$f_Y(x, o) \overset{(1)}{\neq} f_Y(x', o) = f_Y(x', z, o_{-z}) \overset{(2)}{=} f_Y(x, z, o_{-z}) = f_Y(x, o).$$

□

Plan

The evaluation time

Imagine a country where the parliament votes on bills: if a bill passes, it is signed into law on January 1st of the next year.

Suppose it is September. The parliament has just voted on a bill which failed by one vote. Alice did not vote for the bill. Uttered in September:

The evaluation time

Imagine a country where the parliament votes on bills: if a bill passes, it is signed into law on January 1st of the next year.

Suppose it is September. The parliament has just voted on a bill which failed by one vote. Alice did not vote for the bill. Uttered in September:

(5) a. If Alice had voted for the bill, it would be law.
    b. If Alice had voted for the bill, it would become law.

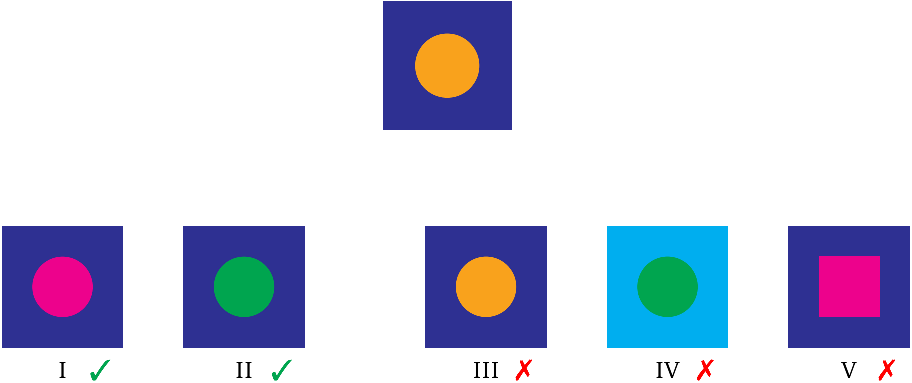
We already have many sophisticated theories of the semantics of tense and tense-modal interaction (Condoravdi 2002). We face a choice:

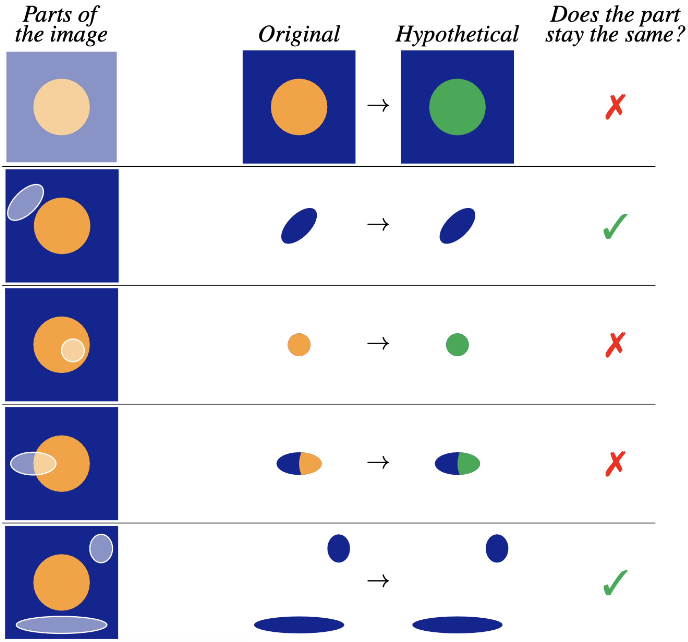
1. Reconstruct existing work on semantics of tense in structural causal models, or
2. Translate structural causal models into the existing frameworks for representing tense.

Plan

A semantics of conditionals (McHugh 2022, 2023)

1. Pick a time at which to imagine the change.
    - This is the *intervention time* t.
2. Vary the part of the world the antecedent A is about at intervention time.
    - This gives us a set of time slices, called the *A-variants of w at t*.
3. Play the laws forward.
    - Find the lawful futures of the A-variants of w at t.
4. Stick on the actual past.
    - This gives us the *modal horizon of A at w*.
5. Restrict to those worlds where the antecedent is true.
6. Check whether the consequent is true at the resulting world(s).


I ✓
II ✓
III ✗
IV ✗
V ✗


Parts of the image
Original
Hypothetical
Does the part stay the same?
✗
✓
✗
✗
✓

Some parts that change.

Some parts that stay the same.

Some parts that change.

Some parts that stay the same.

- We are asked to change the circle.

Some parts that change.

Some parts that stay the same.

- We are asked to change the circle.
- A part stays the same just in case it does not overlap the circle.

Some parts that change.

Some parts that stay the same.

- We are asked to change the circle.
- A part stays the same just in case it does not overlap the circle.
- Two parts overlap just in case they have a part in common.

If x and y are two individuals, then their mereological difference,

$$x - y$$

is the largest individual contained in x which has no part in common with y. (Simons 1987, p. 14, going back to Leśniewski 1927–1931; see Sinisi 1983, 29, Definition VII)

If x and y are two individuals, then their mereological difference,

$$x - y$$

is the largest individual contained in x which has no part in common with y. (Simons 1987, p. 14, going back to Leśniewski 1927–1931; see Sinisi 1983, 29, Definition VII)

If x and y are two individuals, then their mereological difference,

$$x - y$$

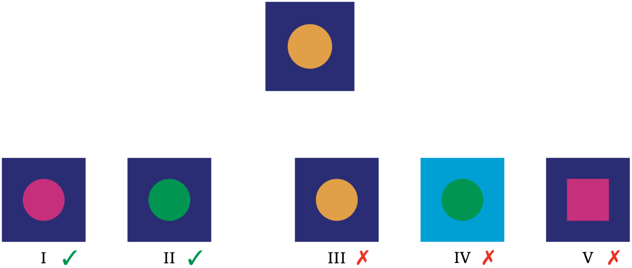
is the largest individual contained in x which has no part in common with y. (Simons 1987, p. 14, going back to Leśniewski 1927–1931; see Sinisi 1983, 29, Definition VII)

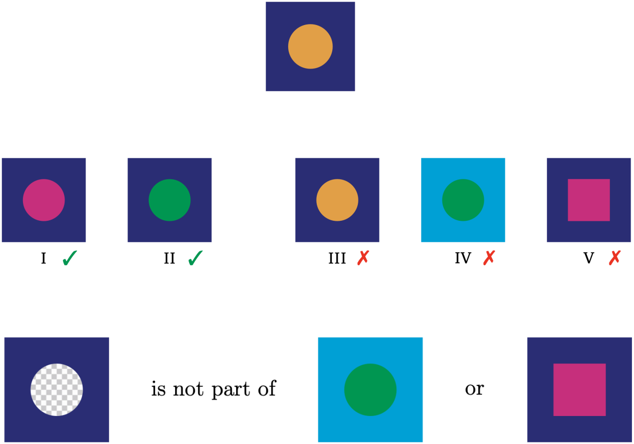
If x and y are two individuals, then their mereological difference,

$$x - y$$

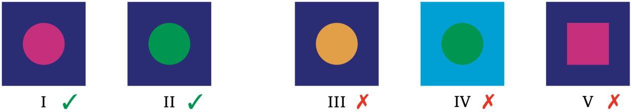
is the largest individual contained in x which has no part in common with y. (Simons 1987, p. 14, going back to Leśniewski 1927–1931; see Sinisi 1983, 29, Definition VII)

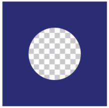
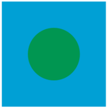
$$\underbrace{x - y}_{ceteris} \; \underbrace{\leq \; z}_{paribus}$$


I ✓
II ✓
III ✗
IV ✗
V ✗



I ✓
II ✓
III ✗
IV ✗
V ✗
is not part of
or



 is not part of  or 

“The circle is a different colour” is not true at 

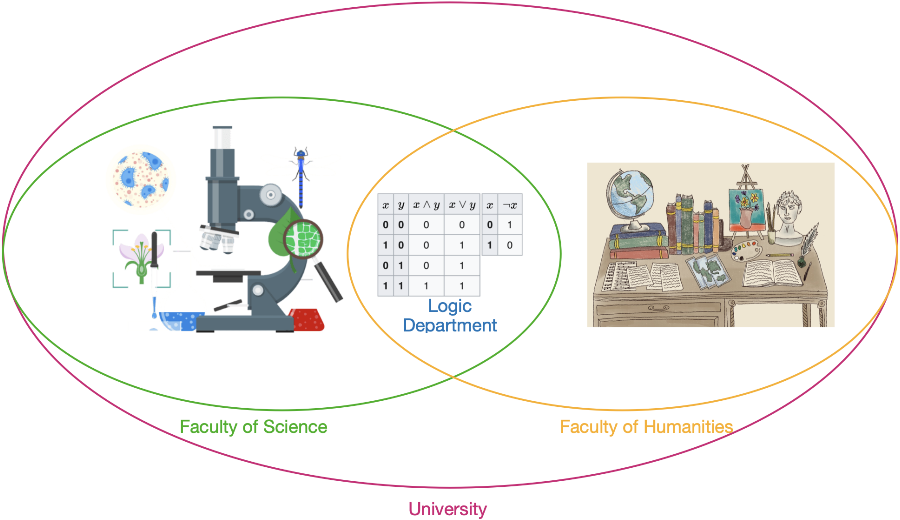
- **The foreground:** the set of states A is about.
- **The background:** the set of states that do not overlap a state in the foreground.

- **The foreground:** the set of states A is about.
- **The background:** the set of states that do not overlap a state in the foreground.

Ceteris paribus

- The background is the *ceteris*, the ‘all else’ in ‘all else being equal’
- *Paribus* means having the *ceteris* as part

Parthood in conceptual space

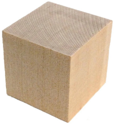


x	y	$x \wedge y$	$x \vee y$
0	0	0	0
1	0	0	1
0	1	0	1
1	1	1	1

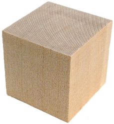
x	$\neg x$
0	1
1	0

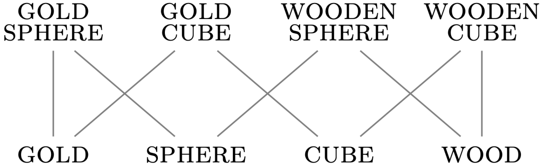
Parthood in conceptual space

Parthood in conceptual space



Parthood in conceptual space





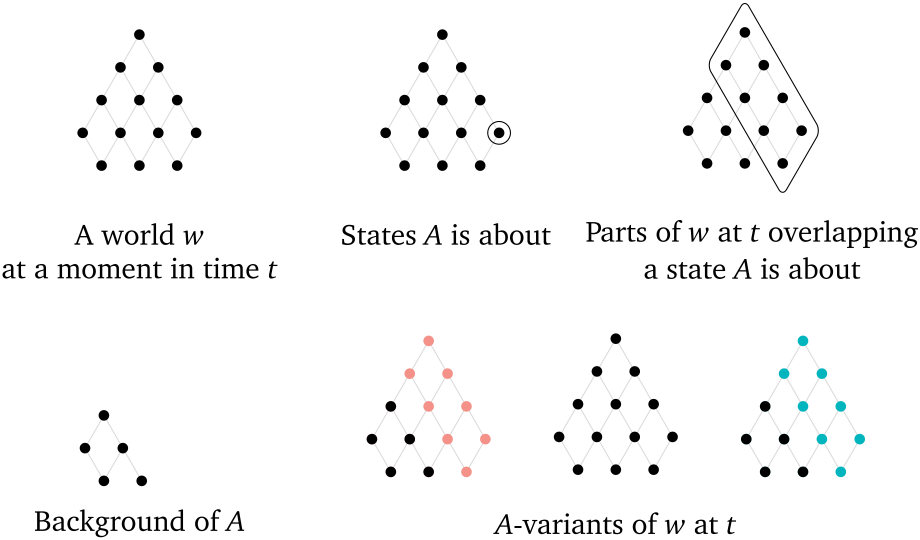


Figure: Steps to construct the A-variants of a world at a moment in time.

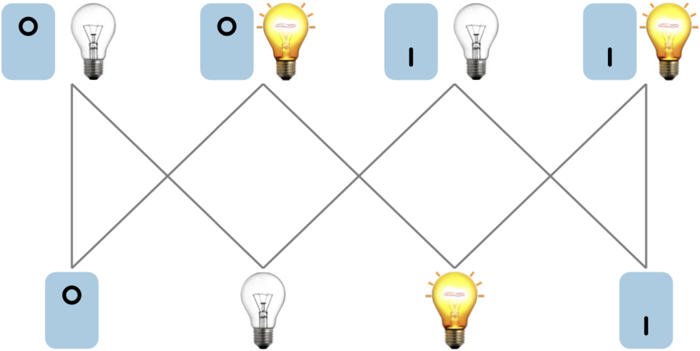


Figure: A state space of the switch and light.

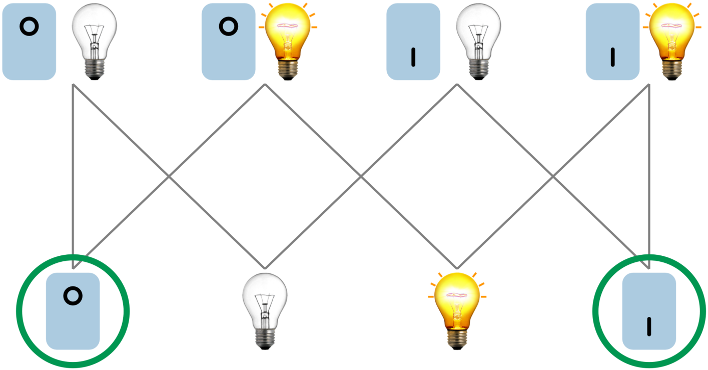


Figure: The states that “the switch is up” is about.

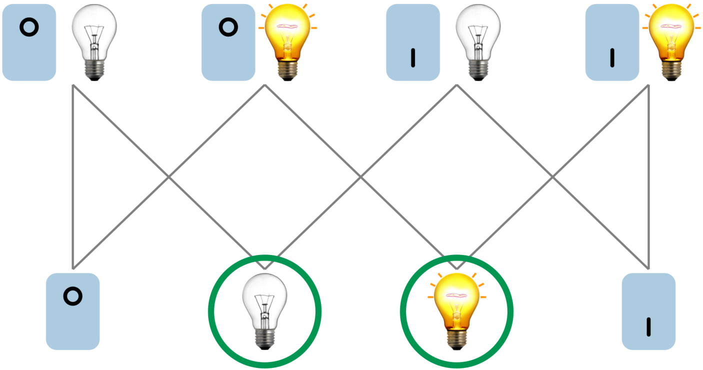


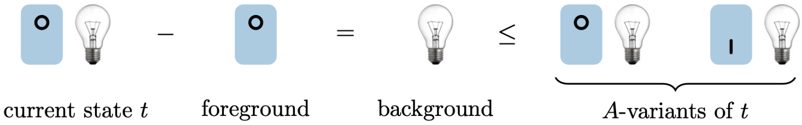
Figure: The states that “the light is off” is about.

(6) *Current state: the switch is up and the light off.*

a. If the switch were down, the light would be on.

b. The light is off because the switch is up.

"The switch is down" and "The switch is up" are about the state of the switch, and not about the state of the light.



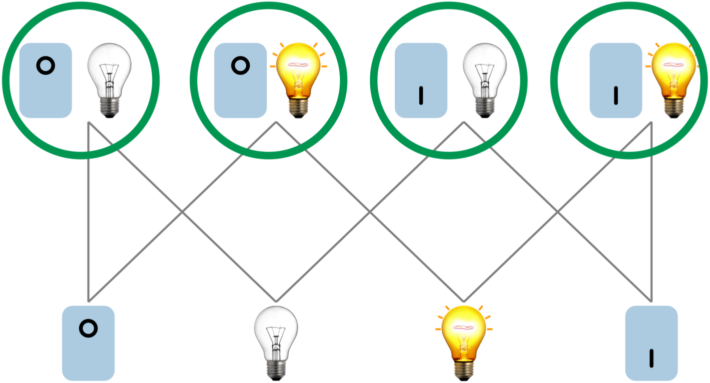
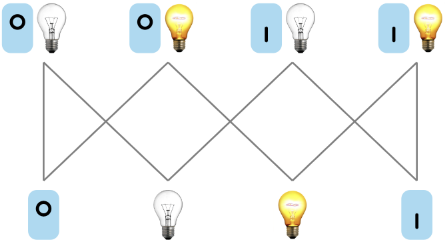
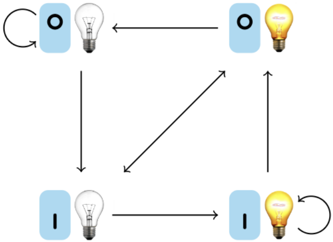


Figure: The states that "the switch is up and the light is off" is about.



(a) Mereological structure.



(b) Nomic possibilities.

Figure: Light switch example. Nomically possible worlds correspond to directed paths in (b).

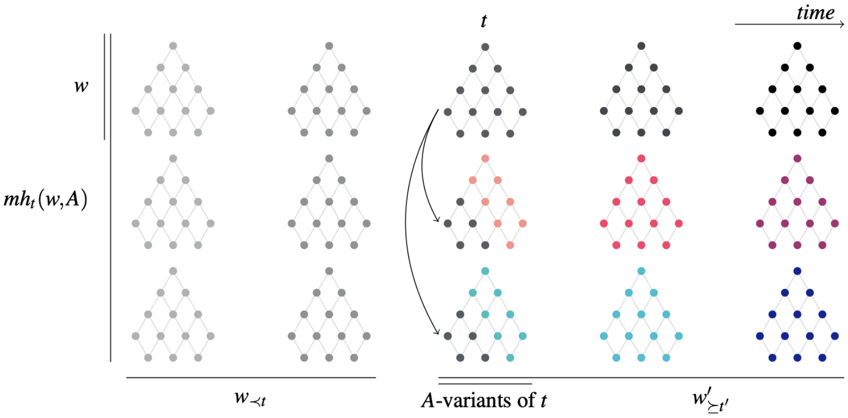


Figure: Constructing the modal horizon.

Definition (Nomic aboutness model)

Where S is a set and $\leq$ a binary relation on S, define

$$\begin{aligned} Sit &:= S \times I, \text{ where } I \text{ is an arbitrary label set,} \\ M &:= \{t_i \in Sit : t \leq u \text{ implies } t = u \text{ for all } u \in S\}, \\ W &:= \{(M', \preceq) : M' \subseteq M, \preceq \text{ is a linear order}\}. \end{aligned}$$

Definition (The modal horizon)

For any sentence A, moment $t \in M$ and world $w \in W$, define

$$mh_{P,t}(w, A) := \{w_{\prec t} \frown w'_{\succeq t'} : t' \text{ is an } A\text{-variant of } t, t' \in w' \text{ and } w' \in P\}.$$

(7) Where P is the set of nomically possible worlds, t the intervention time, and s the selection function,

$$\begin{aligned} A \gg C \text{ is true at } w &\quad \text{iff} \quad mh_{P,t}(w, A) \cap |A| \subseteq |C| \\ A > C \text{ is true at } w &\quad \text{iff} \quad s\big(w, mh_{P,t}(w, A) \cap |A|\big) \in |C| \end{aligned}$$

Plan

The dynamics

- Let $M = (U, V, F, R)$ be a structural causal model.

The dynamics

- Let $M = (U, V, F, R)$ be a structural causal model.
- Define that a *state* of M is an assignment of values to the variables; that is, a function $s : U \cup V \to R(U \cup V)$.

The dynamics

- Let $M = (U, V, F, R)$ be a structural causal model.
- Define that a *state* of M is an assignment of values to the variables; that is, a function $s : U \cup V \to R(U \cup V)$.
- A *path* of M is a sequence of states of M, $(s_0, s_1, \dots)$.

The dynamics

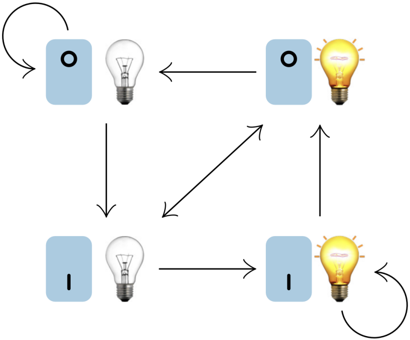
- Let $M = (U, V, F, R)$ be a structural causal model.
- Define that a *state* of M is an assignment of values to the variables; that is, a function $s : U \cup V \to R(U \cup V)$.
- A *path* of M is a sequence of states of M, $(s_0, s_1, \dots)$.
- A path of M is *lawful* (or, *nomically possible*) just in case the states transition according to the structural equations:

$$s_{t+1}(X) = f_X(s_t(PA_X))$$

  for all $t \in \mathbb{N}$ and $X \in V$.

$$L = S$$

(a) Structural causal model



(b) Its dynamic interpretation

$$s_{t+1}(X) = f_X(s_t(PA_X))$$

Figure: A lawful world.

$$s_{t+1}(X) = f_X(s_t(PA_X))$$

Figure: A lawful world.

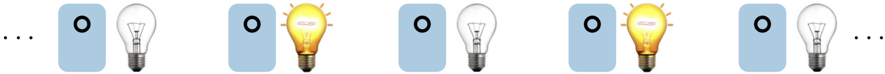
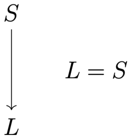


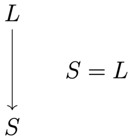
Figure: An unlawful world.

It is unlawful because, e.g. $s_1(L) = 1$ but $f_L(s_t(PA_L)) = f_L(s_t(S)) = f_L(0) = 0$.

Plan



(a) ✓



(b) ✗

Figure: How do we decide that model (a) is correct and model (b) is incorrect?

Figure: Lawful according to (a), unlawful according to (b).

- We can prove that model (b) is incorrect by showing that this world is in fact lawful.
- Actuality implies possibility: we can prove that it is lawful by proving that it is actual—by flicking the switch and seeing the light turn on.

Plan

$$Y = X$$
$$Z = Y$$

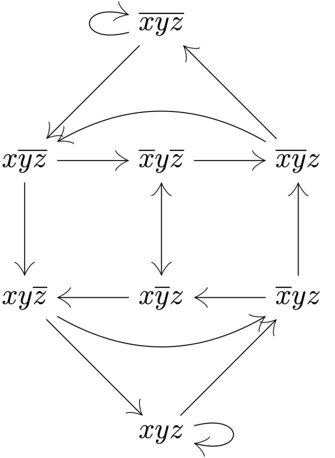
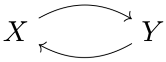


Figure: The dynamic interpretation of a chain.



$$Y = X$$
$$X = Y$$

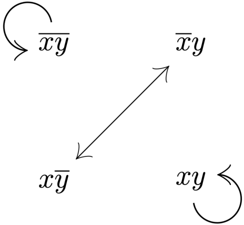
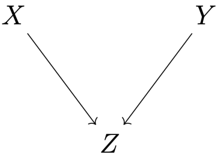


Figure: The dynamic interpretation of a cycle.



$$Z = X \wedge Y$$

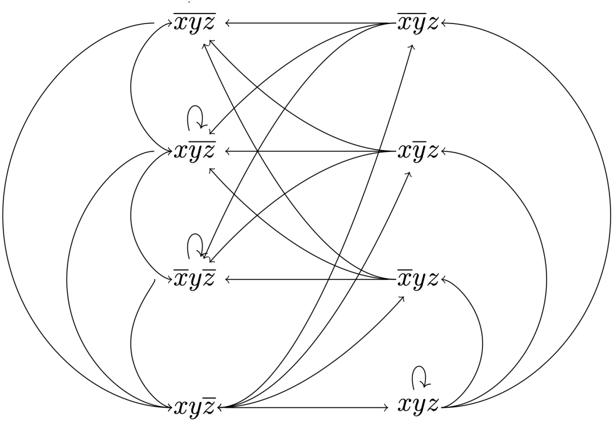


Figure: The dynamic interpretation of an AND-gate.

Plan

Previous hints at the dynamic interpretation

> *The choice of PA_i (connoting parents) is not arbitrary, but expresses the modeller's understanding of which variables Nature must consult* ***before*** *deciding the value of V_i.*
>
> *(Pearl 2009, 203, note 3, my emphasis).*

The independence of X and Y in the graph $X \rightarrow Z \leftarrow Y$

> *reflects our understanding of how causation operates in time; events that are independent in the present do not become dependent merely because they may have common effects in the future.*
>
> *(Pearl, Glymour, and Jewell 2016, p. 41).*

Previous hints at the dynamic interpretation

In temporal metaphors, this three-step procedure [for evaluating counterfactuals] can be interpreted as follows. Step 1 explains the past (U) in light of the current evidence e; step 2 bends the course of history (minimally) to comply with the hypothetical condition $X = x$*; finally, step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition,* $X = x$*.*

(Pearl 2009, p. 37)

Plan

Fixing the exogenous variables

$$Y = X$$
$$Z = Y$$

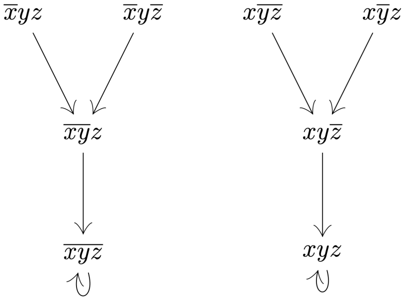
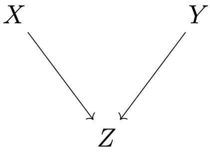


Figure: The dynamic interpretation of a chain, with the exogenous variables fixed.

Fixing the exogenous variables



$$Z = X \wedge Y$$

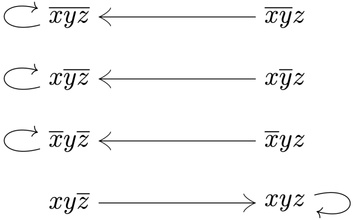


Figure: The dynamic interpretation of an AND-gate, with the exogenous variables fixed.

It is one of the consolations of philosophy that the benefit of showing how to dispense with a concept does not hinge on dispensing with it.

(Quine 1960, p. 189)

Fixing the exogenous variables

A path of M is lawful, while fixing the exogenous variables, just in case

1. it respects the structural equations, interpreted dynamically

$$s_{t+1}(X) = f_X(s_t(PA_X))$$

   for all $t \in \mathbb{N}$ and $X \in V$, and

2. the exogenous variables do not change

$$s_{t+1}(X) = s_t(X)$$

   for all $t \in \mathbb{N}$ and $X \in U$.

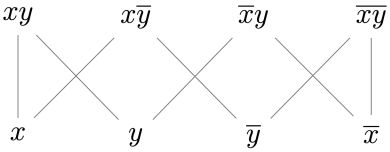
The state space

- A *state* of M is an assignment of values to some of M's variables.
- A state s is *part* of state s' just in case s assigns the same value as s' to all of the variables that receive a value in s.

$X \qquad Y$

$$R(X) = \{x, \overline{x}\}$$
$$R(Y) = \{y, \overline{y}\}$$

(a) A variable set $V = \{X, Y\}$.



(b) The variable space generated by V.

Figure: Translating variables into a state space

$$S_M = \{s : \vec{Y} \to R(\vec{Y}) \mid \vec{Y} \text{ is a nonempty subset of } U \cup V\}$$
$$\leq_M = \subseteq$$

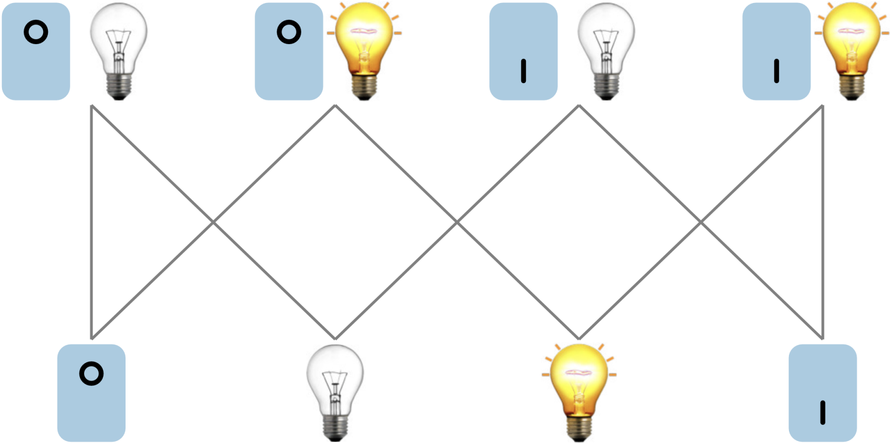


Figure: A state space of the switch and light.

Aboutness

- A sentence $\vec{Y} = \vec{y}$ is *about* state s just in case s assigns a value to all and only the variables in $\vec{Y}$.
- Sentence $\vec{Y} = \vec{y}$ is about state s just in case $s : \vec{Y} \to R(\vec{Y})$.

$$\mathcal{A}_M = \left\{ (\vec{Y} = \vec{y}, s) \mid s : \vec{Y} \to R(\vec{Y}) \right\}$$

The interpretation function

$$|Y = y|_M = \left\{(s_0, s_1, \dots) \in worlds(M) : \text{for some } t, s_{t'}(Y) = y \text{ for all } t' \geq t\right\}$$

Theorem (Theorem 1.6.1, McHugh 2023)

Let M be a recursive structural causal model, $\vec{u}$ a setting of the exogenous variables, $\vec{X} = \vec{x}$ an assignment of values to some variables and φ a Boolean combination of assignments of values to variables. Let

$$M' = (S_M, \leq_M, \mathcal{A}_M, |\cdot|_M, P_{M,\vec{X}=\vec{x}})$$

be defined as above. Let w be any world of M' and t any moment of w such that $t(\vec{U}) = \vec{u}$. Then

$$M, \vec{u} \models [\vec{X} = \vec{x}]\varphi \qquad \text{if and only if} \qquad M', w, t \models \vec{X} = \vec{x} \gg \varphi.$$

References I

Condoravdi, Cleo (2002). Temporal Interpretation of Modals: Modals for the Present and for the Past. *The Construction of Meaning*. CSLI Publications, pp. 59–88. URL: https://semanticsarchive.net/Archive/2JmZTIwO/temp-modals.ps.

Cook, Thomas D. and Donald T. Campbell (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin Boston.

Gasking, Douglas (1955). Causation and recipes. *Mind* 64.256, pp. 479–487. DOI: 10.1093/mind/LXIV.256.479.

Hájek, Alan (2023). Interpretations of Probability. *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford University.

Hall, Ned (2000). Causation and the Price of Transitivity. *Journal of Philosophy* 97.4, pp. 198–222. DOI: 10.2307/2678390.

References II

- Halpern, Joseph Y (2016). *Actual Causality*. MIT Press.
- Halpern, Joseph Y and Judea Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British journal for the philosophy of science* 56.4, pp. 843–887. DOI: 10.1093/bjps/axi147.
- Holland, Paul W (1986). Statistics and causal inference. *Journal of the American statistical Association* 81.396, pp. 945–960.
- Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung,* Ergebnisse Der Mathematik; translated as *Foundations of Probability*. Chelsea Publishing Company, 1950.
- Leśniewski, Stanisław (1927–1931). O podstawach matematyki [On the Foundations of Mathematics], I–V. *Przeglad Filozoficzny*. Volume 30 (1927), 164–206; 31 (1928), 261–291; 32 (1929), 60–101; 33 (1930), 77–105; 34 (1931), 142–170. Abridged English translation by Vito Sinisi 1983.

References III

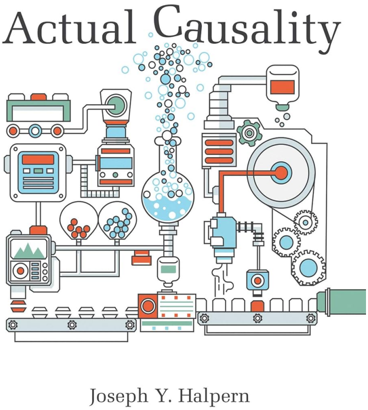
- McHugh, Dean (2022). Aboutness and Modality. *Proceedings of the 23rd Amsterdam Colloquium*, pp. 194–206. DOI: 10.21942/uva.21739718.
- — (2023). Causation and Modality: Models and Meanings. PhD thesis. University of Amsterdam. URL: https://eprints.illc.uva.nl/id/eprint/2243.
- Menzies, Peter and Huw Price (1993). Causation as a secondary quality. *The British Journal for the Philosophy of Science* 44.2, pp. 187–203. DOI: 10.1093/bjps/44.2.187.
- Pearl, Judea (2000). *Causality*. Cambridge University Press.
- — (2009). *Causality, second edition*. Cambridge University Press.
- Pearl, Judea, Madelyn Glymour, and Nicholas P Jewell (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.
- Quine, Willard Van Orman (1960). *Word and object*. MIT press.

References IV

Simons, Peter (1987). *Parts: A study in ontology*. Oxford University Press.

Sinisi, Vito (1983). Leśniewski's foundations of mathematics. *Topoi* 3.1, pp. 3–52. DOI: 10.1007/BF00139700.

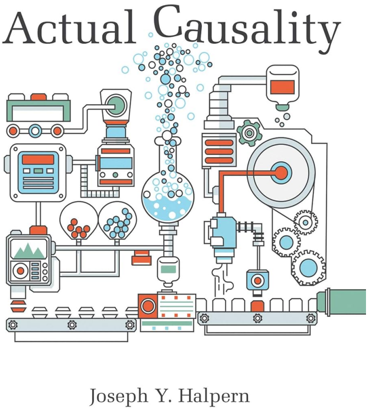
Woodward, James (2016). Causation and Manipulability. *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University.



Halpern (2016), *Actual Causality*:

C is an actual cause of E just in case

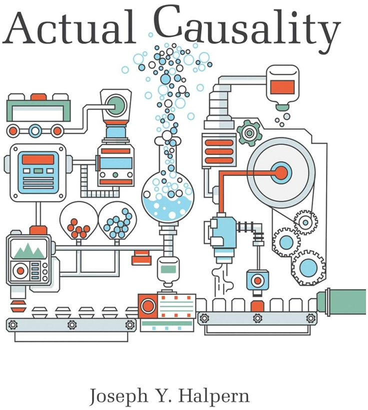
1. C and E actually occurred.



Halpern (2016), *Actual Causality*:

C is an actual cause of E just in case

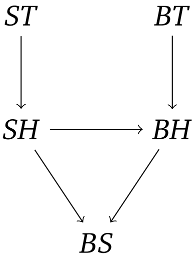
1. C and E actually occurred.
2. There is a set of variables such that, holding them fixed at their actual values, if the cause had not occurred, the effect would not have occurred.



Halpern (2016), *Actual Causality*:

C is an actual cause of E just in case

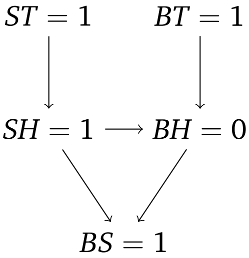
1. C and E actually occurred.
2. There is a set of variables such that, holding them fixed at their actual values, if the cause had not occurred, the effect would not have occurred.
3. C is minimal: no proper subset of C satisfies (1) and (2).



$$SH = ST$$
$$BH = BT \wedge \neg SH$$
$$BS = SH \vee BH$$

Figure: Halpern's model of the Billy and Suzy case (2016, p. 31)

Halpern's account of the Billy and Suzy case



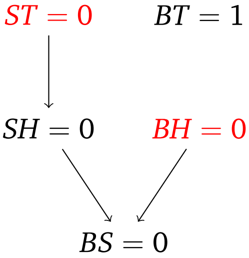
$$SH = ST$$

$$BH = BT \wedge \neg SH$$

$$BS = SH \vee BH$$

Figure: Halpern's model of *Late preemption* (2016, p. 31)

Halpern's account of the Billy and Suzy case



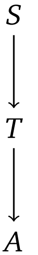
$SH = ST$

$BH = 0$

$BS = SH \vee BH$

Figure: Halpern's model of *Late preemption* (2016, p. 31)

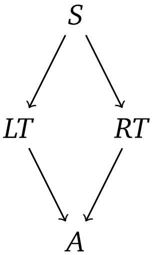
Two models of the switching scenario



$$T = \begin{cases} L & \text{if } S \\ R & \text{if } \neg S \end{cases}$$

$$A = \begin{cases} 1 & \text{if } T = L \vee T = R \\ 0 & \text{otherwise} \end{cases}$$

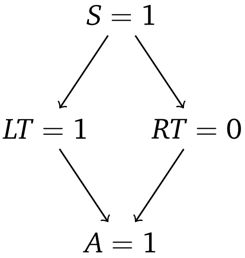
(a) One-variable model



$$LT = S$$
$$RT = \neg S$$
$$A = LT \vee RT$$

(b) Two-variable model

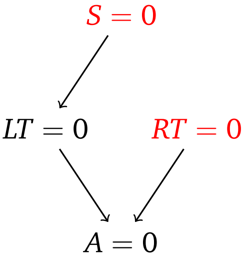
The two-variable model



$$LT = S$$
$$RT = \neg S$$
$$A = LT \vee RT$$

Figure: Two-variable model

The two-variable model



$$LT = S$$
$$RT = 0$$
$$A = LT \vee RT$$

Figure: Two-variable model

Comparing the two models, Halpern and Pearl (2005, p. 872) write:

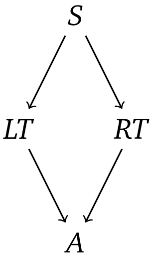
> *The two-variable model depicts the tracks as two independent mechanisms, thus allowing one track to be set (by action or mishap) to false (or true) without affecting the other. Specifically, this permits the disastrous mishap of flipping the switch while the left track is malfunctioning. More formally, it allows a setting where* $S = 1$ *and* $RT = 0$*. Such abnormal settings are imaginable and expressible in the two-variable model, but not in the one-variable model.*

The two-variable model also allows a setting where $S = 0$ and $RT = 0$. The one-variable model rules this out as part of its variable structure.

$$T = \begin{cases} L & \text{if } S \\ R & \text{if } \neg S \end{cases}$$

$$A = \begin{cases} 1 & \text{if } T = L \vee T = R \\ 0 & \text{otherwise} \end{cases}$$

(a) One-variable model



$$LT = S$$
$$RT = \neg S$$
$$A = LT \vee RT$$

(b) Two-variable model

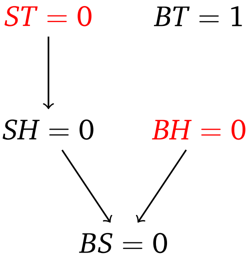
Figure: Two models of the switching scenario

In the two-variable model, one can intervene to make

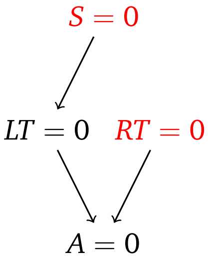
$$S = 0, LT = 0 \text{ and } RT = 0.$$

That is, interventions can make train disappear from the tracks!

The two-variable model

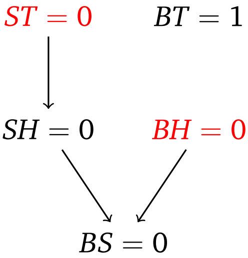


(a) Witness to Suzy causing the window to break

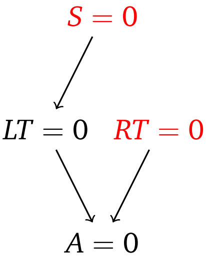


(b) Witness to the switch causing the train to arrive

The two-variable model



(a) Witness to Suzy causing the window to break



(b) Witness to the switch causing the train to arrive

If Billy's rock can disappear mid-flight,
why can't the train disappear mid-journey as well?

Comparing the two models, Halpern and Pearl (2005, p. 872) write:

> *The two-variable model depicts the tracks as two independent mechanisms, thus allowing one track to be set (by action or mishap) to false (or true) without affecting the other. Specifically, this permits the disastrous mishap of flipping the switch while the left track is malfunctioning. More formally, it allows a setting where $S = 1$ and $RT = 0$. Such abnormal settings are imaginable and expressible in the two-variable model, but not in the one-variable model.*

Questions

Is Halpern's solution to the Billy and Suzy case too sensitive to the choice of model?

Which solution do you prefer? Using production + Sartorio's Principle, or Halpern's solution?