Unravelling Sartorio's Difference-making Principle

Dean McHugh

Institute of Logic, Language and Computation
University of Amsterdam

Friday, 22 September 2023
TbILLC 2023, Telavi, Georgia

Institute for Logic,
Language and Computation

University
of Amsterdam

Plan

Making a difference

Sartorio's analysis of difference-making

Unravelling Sartorio's Principle

The ubiquity of the Perfection Principle

On the pragmatic origins of the Perfection Principle

Plan

Making a difference

Sartorio's analysis of difference-making

Unravelling Sartorio's Principle

The ubiquity of the Perfection Principle

On the pragmatic origins of the Perfection Principle

(1) a. Drinking a lot of Georgian wine **caused** me to swim in the pool last night.
b. I am dehydrated **because** I ate too much salty cheese.

(2) a. The employer did not hire Elisabeth Dekker **because** she is pregnant.
b. ChatGPT being trained on far-right Reddit posts **caused** it to output racist stereotypes.

CAUSATION AND MODALITY

Dean McHugh

The modelling question

What information do we use when we judge that a causal claim holds? In other words, what information should a causal model contain?

The meaning question

Under what conditions is a causal claim true or false? What do causal claims mean?

"We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it."

(Lewis 1973)

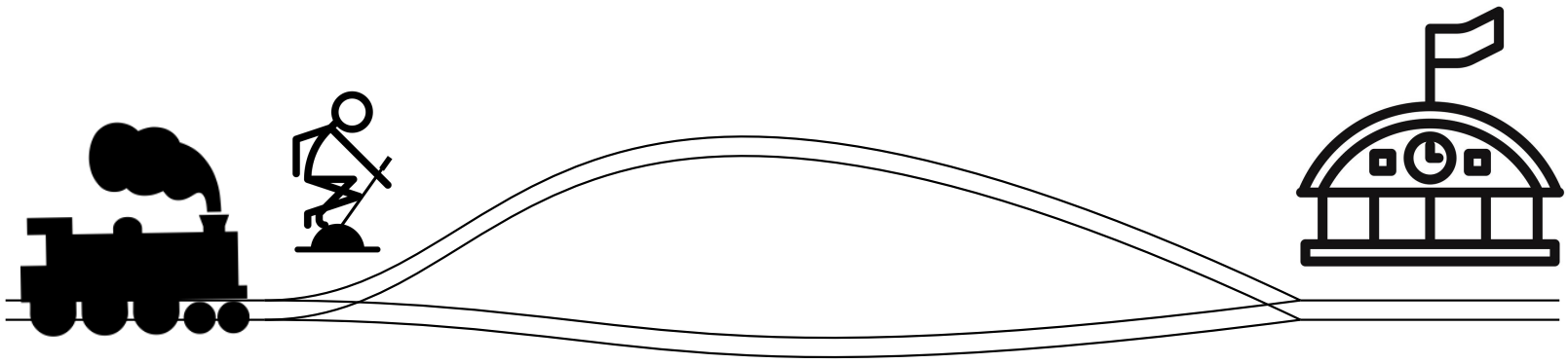


Figure: Switching scenario from Hall (2000, p. 205).

(3) a. The train reached the station because the engineer flipped the switch. ✗

b. The engineer flipping the switch caused the train to reach the station. ✗

C caused E
E because C

$\Rightarrow$

C made a difference to *E*

Today's question

What does "*C* made a difference to *E*" mean?

Hypothesis: Difference making is counterfactual dependence

C made a difference to E

just in case

if C had not occurred, E would not have occurred.

It is not quite clear what 'dependence' is supposed to be, but at least it seems to imply that you would not get the effect without the cause.
The trouble about this is that you might from some other cause. That this effect was produced by this cause does not at all show that it could not, or would not, have been produced by something else in the absence of this cause.

— Elisabeth Anscombe (1971)

Suzy and Billy, expert rock-throwers, are engaged in a competition to see who can shatter a target window first.

They both pick up rocks and throw them at the window, but Suzy throws hers before Billy. Consequently Suzy's rock gets there first, shattering the window.

Since both throws are perfectly accurate, Billy's would have shattered the window if Suzy's had not occurred.

(Hall 2004, p. 235)

(4) The window broke because Suzy threw her rock at it. ✓

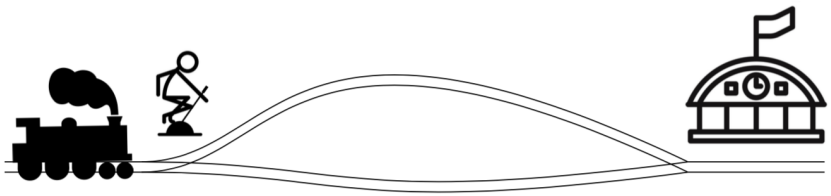
(5) Suzy throwing her rock at the window caused it to break. ✓

Hypothesis: Difference making is counterfactual dependence

C made a difference to E

just in case

if C had not occurred, E would not have occurred.



One thing that catches the eye ... is that, just as the *flip* doesn't make a difference to the [train reaching the station], the *failure to flip* wouldn't have made a difference to the [train reaching the station] either. In other words, *whether or not* I flip the switch makes no difference [to the train's arrival], it only helps to determine the route that the train takes [to the station].

(Sartorio 2005, pp. 74–75)

Plan

Making a difference

Sartorio's analysis of difference-making

Unravelling Sartorio's Principle

The ubiquity of the Perfection Principle

On the pragmatic origins of the Perfection Principle

Sartorio's Principle

If C caused E, then, had C not occurred, the absence of C wouldn't have caused E.

$$C \textit{ cause } E \quad \Rightarrow \quad \neg C > \neg(\neg C \textit{ cause } E)$$

Sartorio's Principle accounts for the switch case:

- Suppose for reductio that the engineer pulling the lever caused the train to reach the station.
- Then (intuitively) if the engineer hadn't pulled the lever, that would have also caused the train to reach the station.
- Sartorio's principle is violated.
- Then assuming Sartorio's principle, the engineer pulling the lever did not caused the train to reach the station.

Sartorio's Principle also accounts for the Billy and Suzy case:

- Imagine that Suzy had not thrown.
- In that case Billy's rock would have hit the window, and it would have broken anyway.
- Intuitively, Billy's throw caused the window to break.
- But what about Suzy *not* throwing?
  Did that cause the window to break?
  Intuitively not!

Sartorio's Principle

If C caused E, then, had C not occurred, the absence of C wouldn't have caused E.

$$C \textit{ cause } E \quad \Rightarrow \quad \neg C > \neg(\neg C \textit{ cause } E)$$

Sartorio's Principle gives us a principled way to distinguish Suzy and the switch.

- Suzy's throw satisfies Sartorio's Principle.
- Pulling the switch does not.

Sartorio's Principle

If C caused E, then, had C not occurred, the absence of C wouldn't have caused E.

$$C \textit{ cause } E \quad \Rightarrow \quad \neg C > \neg(\neg C \textit{ cause } E)$$

Sartorio's principle is automatically satisfied when the effect counterfactually depends on the cause.

- Suppose $\neg C > \neg E$
- $\neg E$ entails $\neg(\neg C \textit{ cause } E)$
- Hence $\neg C > \neg(\neg C \textit{ cause } E)$

Counterfactual dependence is one way to make a difference.

But, as Suzy shows, it is not the only way to make a difference.

Plan

Making a difference

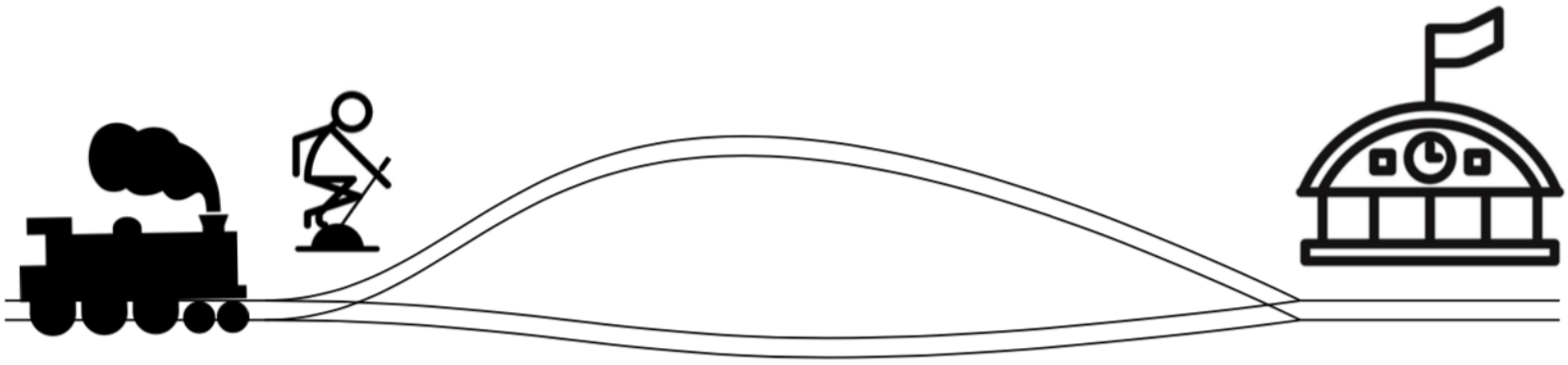
Sartorio's analysis of difference-making

Unravelling Sartorio's Principle

The ubiquity of the Perfection Principle

On the pragmatic origins of the Perfection Principle

Suppose we have a semantics of *cause*, call it *proto-cause*, that does not account for the switches case.



Pulling the lever proto-caused the train to reach the station.

Challenge

How do we amend *proto-cause* to predict that pulling the lever did **not** cause the train to reach the station?

Sartorio's Principle

If C caused E, then, had C not occurred, the absence of C wouldn't have caused E.

$$C \textit{ cause } E \quad \Rightarrow \quad \neg C > \neg(\neg C \textit{ cause } E)$$

$$\begin{aligned} x &\geq 3x - 2 \\ x + 2 &\geq 3x \\ \frac{x+2}{3} - x &\geq 0 \end{aligned}$$

Problem

- The operations of arithmetic have inverses (addition/subtraction; multiplication/division)
- Logical operations

Definition

Let $A[C/B]$ be the result of replacing every occurrence of B in A with C.

Example

$$\big((p \vee q) \wedge \neg q\big)[r/q] \quad = \quad (p \vee r) \wedge \neg r.$$

The Perfection Principle.

For any sentences C and E, there is a sentence X such that C *cause* E entails $C > X$ and $\neg(C > X)[\neg C/C]$.

Sartorio's Principle

$C \textit{ cause } E \;\Rightarrow\; \neg C > \neg(\neg C \textit{ cause } E)$

The Perfection Principle

$C \textit{ cause } E \;\Rightarrow\; (C > X) \;\wedge\; \neg(C > X)[\neg C / C]$

Let $A \diamondsuit\!\!\rightarrow C$ abbreviate $\neg(A > \neg C)$.

(6)
a. **Nonempty domains.** $A > C$ entails $A \diamondsuit\!\!\rightarrow C$.
b. **Stability.** C *cause* E entails $C > (C$ *cause* $E)$.
c. **Idempotence.** $A \diamondsuit\!\!\rightarrow C$ entails $A > (A \diamondsuit\!\!\rightarrow C)$.
d. **Right weakening.**
If C entails C' then $A > C$ entails $A > C'$.
e. If C *cause* E is true, then C is not a subsentence of E.

Theorem

Given the assumptions in (6), Sartorio's Principle is equivalent to the Perfection Principle.

Proof ($\Rightarrow$)

Suppose Sartorio's Principle. Pick any sentences C and E and take $X = (C \textit{ cause } E)$. Then by Stability, $C \textit{ cause } E$ entails $C > X$. We also have the following chain of implications.

$$
\begin{aligned}
C \textit{ cause } E &\Rightarrow \neg C > \neg(\neg C \textit{ cause } E) && \text{(Sartorio's Principle)} \\
&\Rightarrow \neg(\neg C > (\neg C \textit{ cause } E)) && \text{(Nonempty domains)} \\
&\Rightarrow \neg(C > (C \textit{ cause } E))[\neg C/C] \\
& && (C \text{ is not a subsentence of } E) \\
&\Rightarrow \neg(C > X)[\neg C/C] && (X = C \textit{ cause } E)
\end{aligned}
$$

Hence $C \textit{ cause } E$ entails $C > X$ and $\neg(C > X)[\neg C/C]$.

Proof ($\Leftarrow$)

Suppose the Perfection Principle. So $\neg C$ *cause* E entails $(C > X)[\neg C/C]$. Then by contraposition we have (†): $\neg(C > X)[\neg C/C]$ entails $\neg(\neg C$ *cause* $E)$. Observe the following chain of implications.

$$
\begin{array}{lcll}
C \textit{ cause } E & \Rightarrow & \neg(C > X)[\neg C/C] & \text{(Perfection Principle)} \\
 & \Rightarrow & \neg(\neg C > X[\neg C/C]) & \text{(Definition of } [\neg C/C]) \\
 & \Rightarrow & \neg C \diamondsuit\!\!\rightarrow \neg X[\neg C/C] & \text{(Definition of } \diamondsuit\!\!\rightarrow) \\
 & \Rightarrow & \neg C > (\neg C \diamondsuit\!\!\rightarrow \neg X[\neg C/C]) & \text{(Idempotence)} \\
 & \Rightarrow & \neg C > \neg(\neg C > X[\neg C/C]) & \text{(Definition of } \diamondsuit\!\!\rightarrow) \\
 & \Rightarrow & \neg C > \neg(C > X)[\neg C/C] & \text{(Definition of } [\neg C/C]) \\
 & \Rightarrow & \neg C > \neg(\neg C \textit{ cause } E) & (\dagger \text{ and right weakening})
\end{array}
$$

- Presumably C *proto-cause* E entails $C > X$ or $\neg(C > X)[\neg C/C]$, for some X.
- Our theorem gives us a way to turn *proto-cause* into *cause*.
- **Case 1**. If C *proto-cause* E entails $C > X$ then add $\neg(C > X)[\neg C/C]$:

C *cause* E if and only if C *proto-cause* $E \wedge \neg(C > X)[\neg C/C]$

- **Case 2**. If C *proto-cause* E entails $\neg(C > X)[\neg C/C]$ then add $C > X$.

C *cause* E if and only if C *proto-cause* $E \wedge C > X$

Plan

Making a difference

Sartorio's analysis of difference-making

Unravelling Sartorio's Principle

The ubiquity of the Perfection Principle

On the pragmatic origins of the Perfection Principle

- Lewis (1973, p. 536) proposes that an event e causally depends on an event c just in case the following two counterfactuals are true: if c had occurred, e would have occurred.
- Wright (1985, 2011) proposes the NESS (Necessary Element of a Sufficient Set) test for causation, according to which something is a cause just in case there is a set of conditions that are jointly sufficient for the effect, but are not sufficient when the cause is removed from the set.

- Mackie's INUS condition states that a cause is "an insufficient but non-redundant part of a condition which is itself unnecessary but sufficient for the result" (Mackie 1974, p. 64).
- Beckers (2016) use a notion of *production*, arguing that the semantics of *is a cause of* involves comparing the presence and absence of the cause with respect to producing the effect. According to Beckers, C is an cause of E just in case, informally put, C produced E, and after intervening to make $\neg C$ true, $\neg C$ would not have also produced E.

Plan

Making a difference

Sartorio's analysis of difference-making

Unravelling Sartorio's Principle

The ubiquity of the Perfection Principle

On the pragmatic origins of the Perfection Principle

Certainly, it seems to be the case that an inference can, historically, become part of semantic representation in the strict sense; thus, the development of the English conjunction since from a purely temporal word to a marker of causation can be interpreted as a change from a principle of invited inference associated with since (by virtue of its temporal meaning) to a piece of the semantic content of since.

(Geis and Zwicky 1971, pp. 565–566)

Iatridou (1993, 2021) observes that *then* in conditionals takes on a further meaning. She offers the following examples, which are unacceptable with *then* but fine without it.

(7)
a. If I may be frank (*then) you are not looking good today.
b. If John is dead or alive (*then) Bill will find him.
c. If he were the last man on earth (*then) she wouldn't marry him.
d. Even if you give me a million dollars (*then) I will not sell you my piano.

Where $O(p)(q)$ denotes the conditional construction, Iatridou (1993) proposes that *if p then q* asserts $O(p)(q)$ and presupposes $\neg O(\neg p)(q)$.

$$C \ \wedge \ C \gg (C \ \textit{produce} \ E) \ \wedge \ \neg(\neg C \gg (\neg C \ \textit{produce} \ E))$$

(8) C *cause* E and E *because* C entail ...

a. Existential difference-making:
$\neg(\neg C > (\neg C \textit{ produce } E))$

b. Universal difference-making:
$\neg C > \neg(\neg C \textit{ produce } E)$

(9)

a. Reyna was born at Royal Bolton Hospital but received a Danish passport because her mother was born in Copenhagen.

b. He has an American passport because he was born in Boston.

c. I think I was laid off because I'm 56 years old.

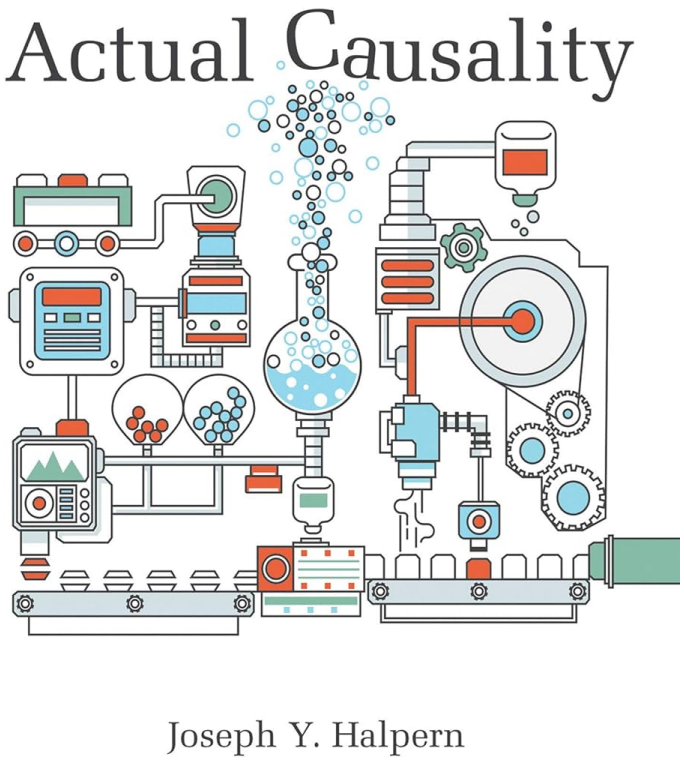
d. Naama Issachar ... could spend up to seven-and-a-half years in a Russian prison because 9.5 grams of cannabis were found in her possession during a routine security check.

e. A 90-day study in 8 adults found that supplementing a standard diet with 1.3 cups (100 grams) of fresh coconut daily caused significant weight loss.

(10) If he hadn't eaten 413 chicken nuggets, he wouldn't have been paralysed.

Summary

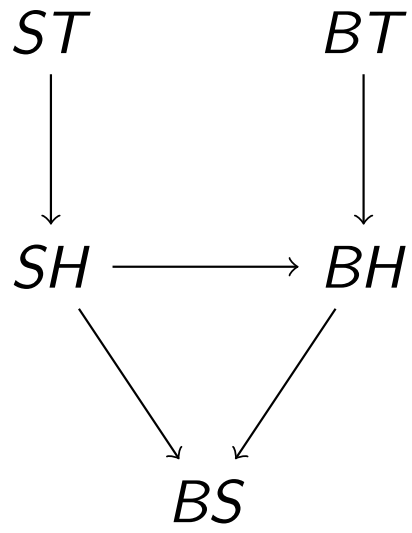
- Sartorio's Principle offers a principled way to distinguish the Billy and Suzy case from switching cases.
- But given the Principle's logical structure, one cannot simply add it to existing semantics of *cause*.
- Our theorem provides a way to add Sartorio's Principle to semantic theories of *cause*.



Halpern (2016), *Actual Causality*:

C is an actual cause of E just in case

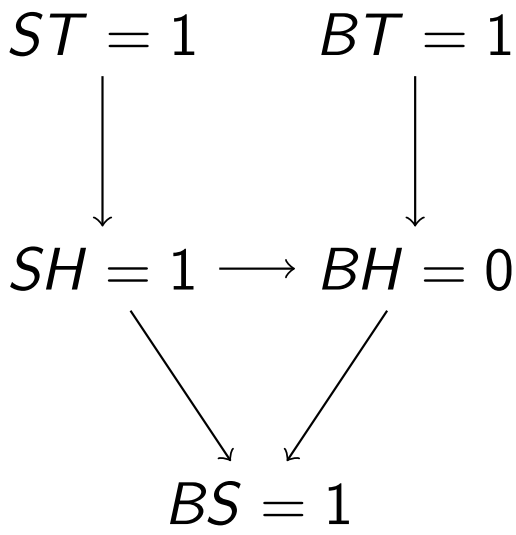
1. C and E actually occurred.
2. There is a set of variables such that, holding them fixed at their actual values, if the cause had not occurred, the effect would not have occurred.
3. C is minimal: no proper subset of C satisfies (1) and (2).



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

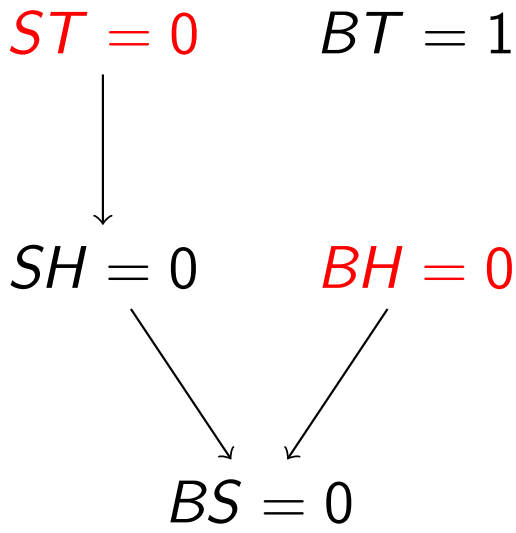
Figure: Halpern's model of the Billy and Suzy case (2016, p. 31)

Halpern's account of the Billy and Suzy case



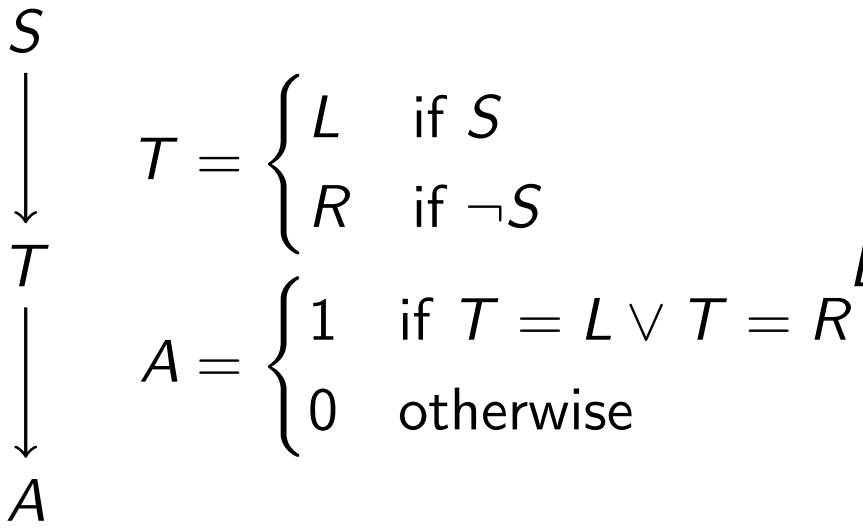
$$SH = ST$$
$$BH = BT \wedge \neg SH$$
$$BS = SH \vee BH$$

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

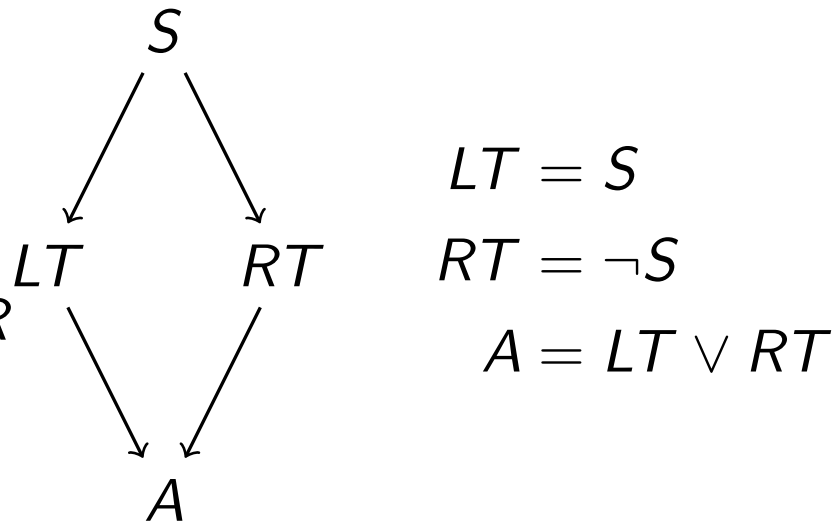


$$SH = ST$$
$$BH = 0$$
$$BS = SH \vee BH$$

Two models of the switching scenario

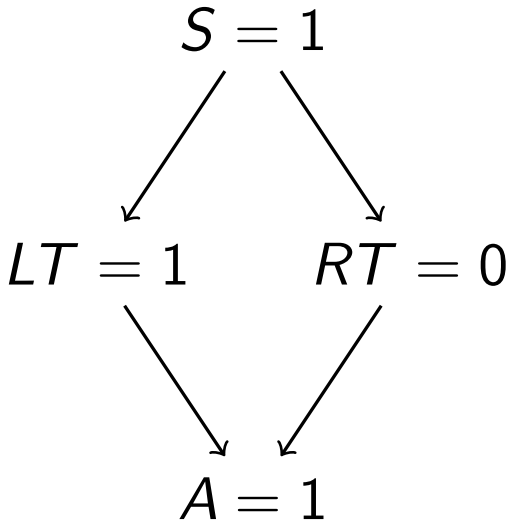


(a) One-variable model



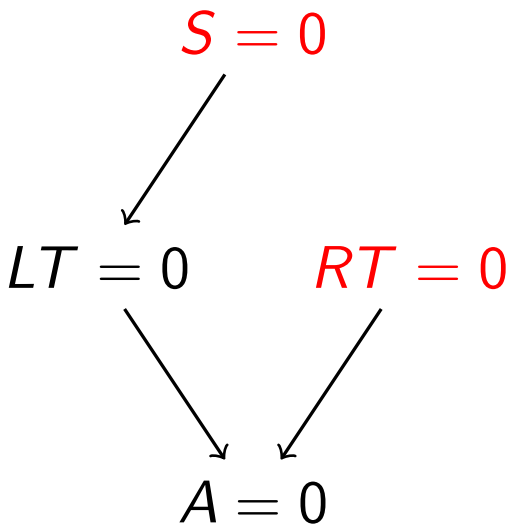
(b) Two-variable model

The two-variable model



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

Figure: Two-variable model

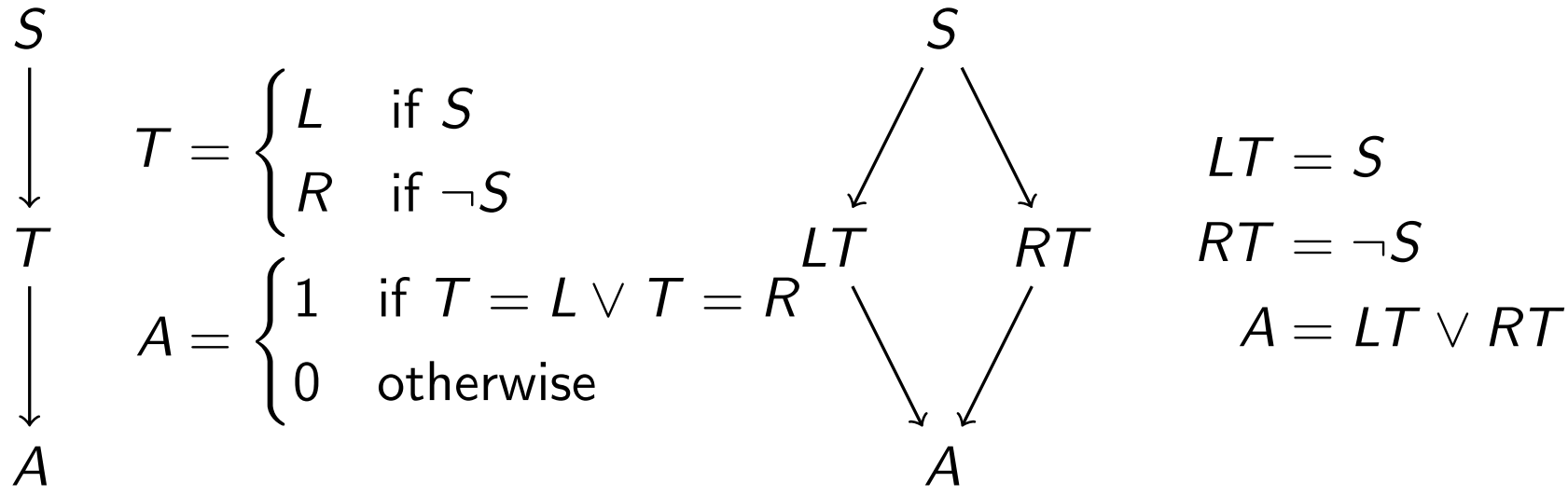


$$LT = S$$
$$RT = 0$$
$$A = LT \vee RT$$

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.



(a) One-variable model

(b) Two-variable model

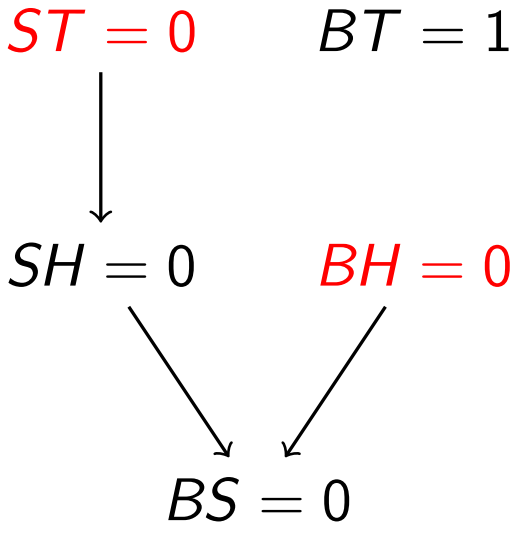
Figure: Two models of the switching scenario

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

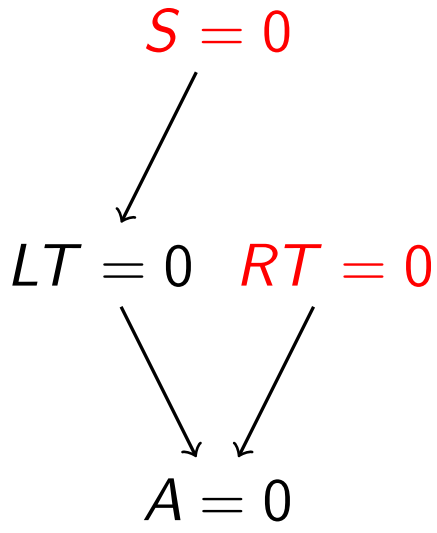
$$S = 0,\ LT = 0 \text{ and } RT = 0.$$

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

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.*

Halpern's solution to the Billy and Suzy case is too sensitive to the choice of model.

Sartorio's Principle offers a more robust solution.

References I

- Anscombe, Gertrude Elizabeth Margaret (1971). *Causality and determination: An inaugural lecture*. CUP Archive.
- Beckers, Sander (2016). Actual Causation: Definitions and Principles. PhD thesis. KU Leuven. URL: https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS1656621&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US.
- Geis, Michael L. and Arnold M. Zwicky (1971). On invited inferences. *Linguistic inquiry* 2.4, pp. 561–566. URL: www.jstor.org/stable/4177664.
- Hall, Ned (2000). Causation and the Price of Transitivity. *Journal of Philosophy* 97.4, pp. 198–222. DOI: 10.2307/2678390.
- — (2004). Two concepts of causation. *Causation and counterfactuals*. Ed. by John Collins, Ned Hall, and Paul Laurie. MIT Press, pp. 225–276.

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.
- Iatridou, Sabine (1993). On the contribution of conditional then. *Natural language semantics* 2.3, pp. 171–199.
- — (2021). Grammar matters. *Conditionals, Paradox, and Probability: Themes from the Philosophy of Dorothy Edgington*. Ed. by Lee Walters and John Hawthorne. Oxford University Press. DOI: 10.1093/oso/9780198712732.003.0008.
- Lewis, David (1973). Causation. *Journal of Philosophy* 70.17, pp. 556–567. DOI: 10.2307/2025310.
- Mackie, John L (1974). *The cement of the universe: A study of causation*. Clarendon Press. DOI: 10.1093/0198246420.001.0001.

References III

- Sartorio, Carolina (2005). Causes As Difference-Makers. *Philosophical Studies* 123.1, pp. 71–96. DOI: 10.1007/s11098-004-5217-y.
- Wright, Richard (1985). Causation in tort law. *California Law Review* 73.6, pp. 1735–1828. DOI: 10.2307/3480373.
- — (2011). The NESS account of natural causation: a response to criticisms. *Perspectives on Causation*. Ed. by Richard Goldberg. Hart Publishing, pp. 13–66.