TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

"Actual Causality"

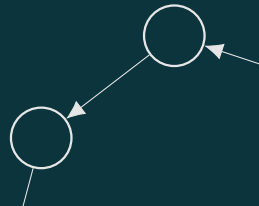
Prof. Dr. Kristian Kersting

Moritz Willig — Today's speaker

Tim Woydt

Florian Busch

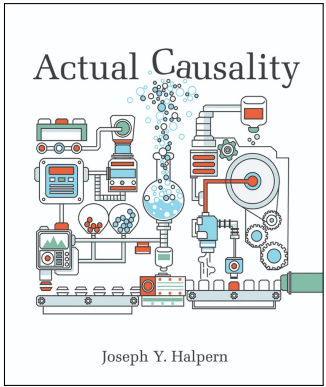
Matej Zečević



Actual Causality

Actual Causality reasons about causes under specific configurations of variables.

Lecture based on the following book:



Book

'*Actual Causality*' by **Joseph Y. Halpern**

Open Access:

https://direct.mit.edu/books/oa-monograph/3451/Actual-Causality

Talk by Joseph Halpern:

https://www.youtube.com/watch?v=hTFs1ploHPw

The Need for Actual Causality

Problem: *do*-calculus tells us if X affects Y *in general*.

- Sometimes called '*type causality*' (or '*general causality*').

The Need for Actual Causality

Problem: *do*-calculus tells us if X affects Y *in general*.

- Sometimes called '*type causality*' (or '*general causality*').
  $\rightarrow$ In general: "smoking causes cancer".
  $\rightarrow$ Can be used for making predictions / forward-looking.

The Need for Actual Causality

Problem: *do*-calculus tells us if X affects Y *in general*.

- Sometimes called '*type causality*' (or '*general causality*').
  $\rightarrow$ In general: "smoking causes cancer".
  $\rightarrow$ Can be used for making predictions / forward-looking.

Commonly does not reason about the *actual* outcome.

$\rightarrow$ 'The fact that Bob smoked for 30 years, did cause his lung cancer.'
$\rightarrow$ 'The fact that a match was dropped, was not responsible for the forest fire.'

The Need for Actual Causality

Problem: *do*-calculus tells us if X affects Y *in general*.

- Sometimes called '*type causality*' (or '*general causality*').
  $\rightarrow$ In general: "smoking causes cancer".
  $\rightarrow$ Can be used for making predictions / forward-looking.

Commonly does not reason about the *actual* outcome.

$\rightarrow$ 'The fact that Bob smoked for 30 years, did cause his lung cancer.'
$\rightarrow$ 'The fact that a match was dropped, was not responsible for the forest fire.'

Actual causality talks about specific instances/scenarios.

- Sometimes called '*token causality*' or '*specific causality*'.
  $\rightarrow$ Token causality is about explanations / backward-looking.

Example: Rock-Throwing

"Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Because both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw."

Lewis, David. "Causation as influence." The Journal of Philosophy 97.4 (2000): 182-197.

Assumptions

Actual causality assumes a fully observed and deterministic world:

- No noise/determinism of the equations.
- Exogenous variables $\mathcal{U}$ are fully specified.

$\rightarrow$ Exogenous variables $\mathcal{U}$ fully determine the system.

Assumptions

Actual causality assumes a fully observed and deterministic world:

- No noise/determinism of the equations.
- Exogenous variables $\mathcal{U}$ are fully specified.

$\rightarrow$ Exogenous variables $\mathcal{U}$ fully determine the system.

Reasons about actual causes within a model.
Not concerned with problem of insufficient knowledge.

Causal Models

A **causal model** is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{F})$.

Causal Models

A **causal model** is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{F})$.

A **signature** is a tuple $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$.

- $\mathcal{U}$: **Exogenous variables** (determined outside the model).
- $\mathcal{V}$: **Endogenous variables** (determined by others variables in the model).
- $\mathcal{R}$: **Associates** with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y.

Causal Models

A **causal model** is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{F})$.

A **signature** is a tuple $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$.

- $\mathcal{U}$: **Exogenous variables** (determined outside the model).
- $\mathcal{V}$: **Endogenous variables** (determined by others variables in the model).
- $\mathcal{R}$: **Associates** with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y.

$\mathcal{F}$ is the set of **structural equations**, where for each $X \in \mathcal{V}$:

$$F_X : \times_{Z \in (\mathcal{U} \cup \mathcal{V} - \{X\})} \mathcal{R}(Y) \to \mathcal{R}(X)$$

Causal Models

A **causal model** is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{F})$.

A **signature** is a tuple $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$.

- $\mathcal{U}$: **Exogenous variables** (determined outside the model).
- $\mathcal{V}$: **Endogenous variables** (determined by others variables in the model).
- $\mathcal{R}$: **Associates** with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y.

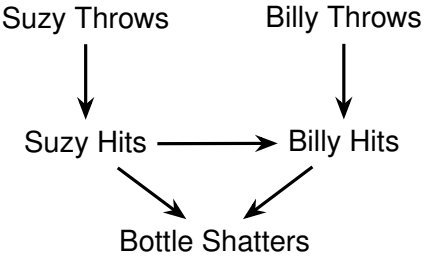
$\mathcal{F}$ is the set of **structural equations**, where for each $X \in \mathcal{V}$:

$$F_X : \times_{Z \in (\mathcal{U} \cup \mathcal{V} - \{X\})} \mathcal{R}(Y) \rightarrow \mathcal{R}(X)$$

A **context** u is a setting for all variables in $\mathcal{U}$
and $(\mathcal{M}, u)$ represents a specific '**world**' where all variables take definite values.

Causal Models - Rock Throwing

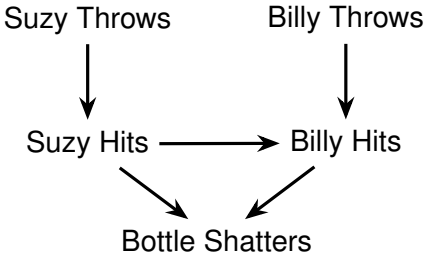
$\mathcal{M} = (\mathcal{S}, \mathcal{F})$



Causal Models - Rock Throwing

$$\mathcal{M} = (\mathcal{S}, \mathcal{F})$$

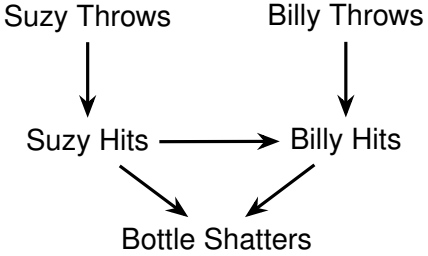
$$\mathcal{S} = \begin{cases} \mathcal{U} & = \{U_{ST}, U_{BT}\} \\ \mathcal{V} & = \{ST, BT, SH, BH, BS\} \\ \mathcal{R}(Y) & = \{0, 1\} \text{ for all } Y \in (\mathcal{V} \cup \mathcal{U}) \end{cases}$$



Causal Models - Rock Throwing

$$\mathcal{M} = (\mathcal{S}, \mathcal{F})$$

$$\mathcal{S} = \begin{cases} \mathcal{U} & = \{U_{ST}, U_{BT}\} \\ \mathcal{V} & = \{ST, BT, SH, BH, BS\} \\ \mathcal{R}(Y) & = \{0, 1\} \text{ for all } Y \in (\mathcal{V} \cup \mathcal{U}) \end{cases}$$



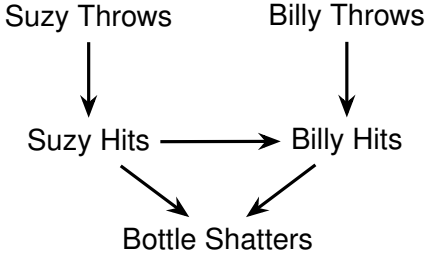
$$\mathcal{F} = \begin{cases} ST & = U_{ST} \\ BT & = U_{BT} \\ SH & = ST \\ BH & = BT \wedge \neg SH \\ BS & = SH \vee BH \end{cases}$$

Causal Models - Rock Throwing

$$\mathcal{M} = (\mathcal{S}, \mathcal{F})$$

$$\mathcal{S} = \begin{cases} \mathcal{U} & = \{U_{ST}, U_{BT}\} \\ \mathcal{V} & = \{ST, BT, SH, BH, BS\} \\ \mathcal{R}(Y) & = \{0, 1\} \text{ for all } Y \in (\mathcal{V} \cup \mathcal{U}) \end{cases}$$

$$\mathcal{F} = \begin{cases} ST & = U_{ST} \\ BT & = U_{BT} \\ SH & = ST \\ BH & = BT \wedge \neg SH \\ BS & = SH \vee BH \end{cases}$$



Actual World:

$u = \{U_{ST} = 1, U_{BT} = 1\}$

Entailment in Causal Models

A given world '$(\mathcal{M}, u)$' with a causal model $\mathcal{M} = (\mathcal{S}, \mathcal{F})$ and a configuration of exogenous variables u entails logical statements:

$$(\mathcal{M}, \mathbf{u}) \models \varphi$$

The language of φ consists of

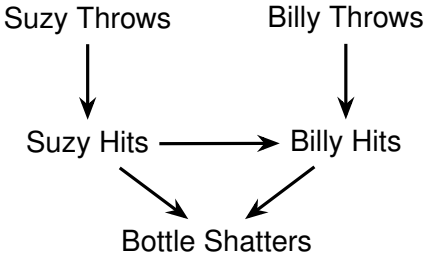
- Primitive events: $X = x$
- 'Interventions': $[X \leftarrow x]\varphi$ ("after setting X to x, φ holds")
- Boolean operations on primitive events.

$$\varphi ::= (X = x) \mid [X \leftarrow x]\varphi \mid (\neg\varphi) \mid (\varphi \wedge \varphi) \mid (\varphi \vee \varphi)$$

Entailment in Causal Models - Rock Throwing

With $\mathcal{M}$ as defined before and actual world $u = \{U_{ST} = 1, U_{BT} = 1\}$:

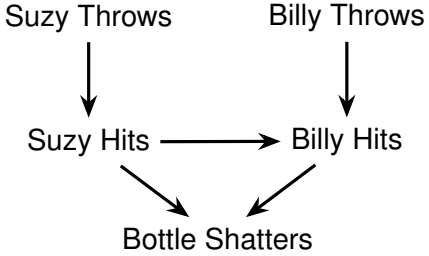
$$(\mathcal{M}, u) \models \underbrace{}_{\text{Actual World}} \underbrace{(SH = 1) \wedge (BT = 1) \wedge (SH = 1) \wedge (BH = 0) \wedge (BS = 1)}_{\text{Entailment}}$$



Entailment in Causal Models - Rock Throwing

With $\mathcal{M}$ as defined before and actual world $u = \{U_{ST} = 1, U_{BT} = 1\}$:

$$(\mathcal{M}, u) \models \underbrace{(SH = 1) \wedge (BT = 1) \wedge}_{\text{Actual World}} \underbrace{(SH = 1) \wedge (BH = 0) \wedge (BS = 1)}_{\text{Entailment}}$$



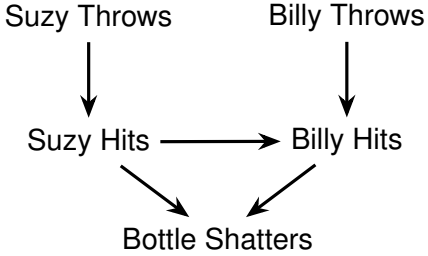
Interventions: $(\mathcal{M}, \{U_{ST} = 0, U_{BT} = 1\}) \models [BH \leftarrow 0]BS = 0$

"Under context u and intervention do($BH = 0$), BS takes value 0 in model $\mathcal{M}$."

Entailment in Causal Models - Rock Throwing

With $\mathcal{M}$ as defined before and actual world $u = \{U_{ST} = 1, U_{BT} = 1\}$:

$$(\mathcal{M}, u) \models \underbrace{(SH = 1) \wedge (BT = 1) \wedge}_{\text{Actual World}} \underbrace{(SH = 1) \wedge (BH = 0) \wedge (BS = 1)}_{\text{Entailment}}$$



Interventions: $(\mathcal{M}, \{U_{ST} = 0, U_{BT} = 1\}) \models [BH \leftarrow 0]BS = 0$

"Under context u and intervention do($BH = 0$), BS takes value 0 in model $\mathcal{M}$."

Intervention Semantics:

$$(\mathcal{M}, u) \models [Y \leftarrow y]\psi \text{ iff } (\mathcal{M}^{Y \leftarrow y}, u) \models \psi$$

‘But-For’ Test

But-for Test: ‘A is a cause of B if, but for A, B would not have happened.’

Preliminary Considerations

1.) A is *necessary* for the occurrence of B.

2.) Saying ‘A causes B’ requires a world where A and B both occurred.
$(\mathcal{M}, u) \models (A = 1) \wedge (B = 1)$

3.) Thinking about A not happening, considers a counterfactual world:
$(\mathcal{M}, u) \models [A \leftarrow 0, \vec{W} \leftarrow \vec{w}](B = 0)$
(...with $\vec{W} \leftarrow \vec{w}$ describing possible additional interventions.)

Rock-Throwing Revisited

> *"Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Because both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw."*

Is 'Suzy throwing' a but-for cause of the bottle shattering?

Rock-Throwing Revisited

> *"Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Because both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw."*

Is 'Suzy throwing' a but-for cause of the bottle shattering?

A naive approach fails:

Even if Suzy would not have hit the bottle, Billy would have shattered the bottle.

$$(\mathcal{M}, \{ST = 1, BT = 1\}) \models [ST \leftarrow 0](BS = 1)$$

$\rightarrow$ For this, Suzy alone would not be a but-for cause of the bottle breaking.

Rock-Throwing Revisited

> *"Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Because both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw."*

Is 'Suzy throwing' a but-for cause of the bottle shattering?

A naive approach fails:

Even if Suzy would not have hit the bottle, Billy would have shattered the bottle.

$$(\mathcal{M}, \{ST = 1, BT = 1\}) \models [ST \leftarrow 0](BS = 1)$$

$\rightarrow$ For this, Suzy alone would not be a but-for cause of the bottle breaking.

But, we *know* that Suzy threw and broke the bottle... how to fix this?

General Requirements for Actual Causes

A primitive event $\vec{X} = \vec{x}$ is an **actual cause** of φ in (M, u) if:

AC1 $(\mathcal{M}, u) \models (\vec{X} = \vec{x})$ and $(\mathcal{M}, u) \models \varphi$.
'The cause and the effect must actually happen.'

AC2 *(see next slides)*

AC3 $\vec{X}$ is minimal.
No subset of $\vec{X}$ satisfies AC1 and AC2.

AC1: Relevant Worlds

$$(\mathcal{M}, u) \models (\vec{X} = \vec{x}) \text{ and } (\mathcal{M}, u) \models \varphi$$

Interpretation: 'Only reason about relevant worlds, where the cause and effect must have actually happened.'

$\neg\vec{X} = \vec{x}$ case: If $\vec{X} = \vec{x}$ did not happen, it can not have caused φ.
$\neg\varphi$ case: If φ did not happen, $\vec{X} = \vec{x}$ can not have caused it.

AC1: Relevant Worlds

$$(\mathcal{M}, u) \models (\vec{X} = \vec{x}) \text{ and } (\mathcal{M}, u) \models \varphi$$

Interpretation: 'Only reason about relevant worlds, where the cause and effect must have actually happened.'

$\neg(\vec{X} = \vec{x})$ case: If $\vec{X} = \vec{x}$ did not happen, it can not have caused φ.
$\neg\varphi$ case: If φ did not happen, $\vec{X} = \vec{x}$ can not have caused it.

To reason about whether $(\vec{X} = \vec{x})$ caused φ, both, $(\vec{X} = \vec{x})$ and φ must have happened.

AC3: Minimality

$\vec{X}$ is minimal.

Interpretation: 'All primitive events in $\vec{X}$ are necessary for the effect.'

If $\vec{X}$ alone is already an actual cause of φ, then adding some unrelated Z to $\vec{X}$ contributes no further to our inference.

$\rightarrow$ If already $\vec{X} = \{A = a\}$ is sufficient to cause φ, then for $\vec{X} = \{A = a, Z = z\}$, it is still A and not Z that causes φ.

AC3: Minimality

$\vec{X}$ is minimal.

Interpretation: 'All primitive events in $\vec{X}$ are necessary for the effect.'

If $\vec{X}$ alone is already an actual cause of φ, then adding some unrelated Z to $\vec{X}$ contributes no further to our inference.

$\rightarrow$ If already $\vec{X} = \{A = a\}$ is sufficient to cause φ, then for $\vec{X} = \{A = a, Z = z\}$, it is still A and not Z that causes φ.

'Suzy throwing' the bottle and the sky being blue, is not a minimal cause, as 'Suzy hitting' would already have been a sufficient cause.

Original Definition of (AC2)

AC2(a) 'Necessity': There exists a partition of $\mathcal{V}$ into disjoint subsets $(\vec{Z}, \vec{W})$, with $\vec{X} \subseteq \vec{Z}$ and a setting $x' \in \vec{X}$ and $\vec{w} \in \vec{W}$ such that:

$$(\mathcal{M}, u) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}] \neg \varphi$$

Interpretation: "*But for the fact that $\vec{X} = \vec{x}'$ occurred, φ would not have happened.*"

As we saw, this definition alone is insufficient!

Original Definition of (AC2)

AC2(a) 'Necessity': There exists a partition of $\mathcal{V}$ into disjoint subsets $(\vec{Z}, \vec{W})$, with $\vec{X} \subseteq \vec{Z}$ and a setting $x' \in \vec{X}$ and $\vec{w} \in \vec{W}$ such that:

$$(\mathcal{M}, u) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varphi$$

Interpretation: "*But for the fact that $\vec{X} = \vec{x}'$ occurred, φ would not have happened.*"
As we saw, this definition alone is insufficient!

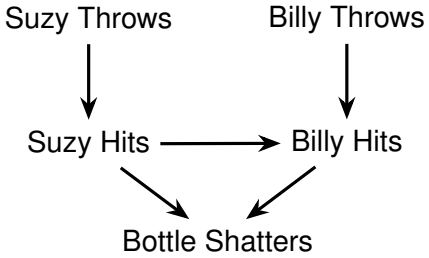
Intermediate conditions $AC2(b^o)$, $AC2(b^u)$ exist that were paired with AC2(a).
$\rightarrow$ Commonly referred to the 'original' (o) and 'updated' (u) definitions.
$\rightarrow$ They are harder to grasp and superseded by the most recent definition.

Original Definition of (AC2)

AC2(a) 'Necessity': There exists a partition of $\mathcal{V}$ into disjoint subsets $(\vec{Z}, \vec{W})$, with $\vec{X} \subseteq \vec{Z}$ and a setting $x' \in \vec{X}$ and $\vec{w} \in \vec{W}$ such that:

$$(\mathcal{M}, u) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varphi$$

$\vec{Z}$ acts as 'causal path' from $\vec{X}$ to the resulting φ.
e.g., $ST \rightarrow SH \rightarrow BS$.

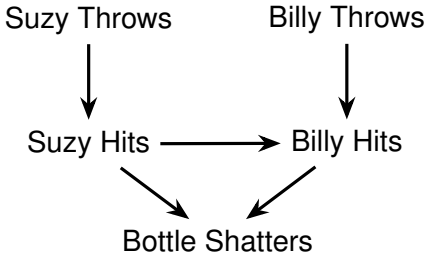


Original Definition of (AC2)

AC2(a) 'Necessity': There exists a partition of $\mathcal{V}$ into disjoint subsets $(\vec{Z}, \vec{W})$, with $\vec{X} \subseteq \vec{Z}$ and a setting $x' \in \vec{X}$ and $\vec{w} \in \vec{W}$ such that:

$$(\mathcal{M}, u) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varphi$$

$\vec{Z}$ acts as 'causal path' from $\vec{X}$ to the resulting φ.
e.g., $ST \rightarrow SH \rightarrow BS$.



$\rightarrow$ To exclude Billy's influence, we must require that some variables $\vec{Z}$ must be held at their *actual* values $\vec{z}^*$.

Example: If a forest fire requires a match (M) and oxygen (O), the original definition might call "Oxygen" the cause because *if* there were no oxygen, the match wouldn't work (even if oxygen is a constant in that environment).

The Need for Modification

Idea: Hold constant values $\vec{z}^*$ along the causal path. **AC2(b^o) 'Sufficiency'**: If $\vec{z}^*$ is such that $(\mathcal{M}, u) \models \vec{Z} = \vec{z}^*$, then for all subsets $\vec{Z}'$ of $\vec{Z} \setminus \vec{X}$, we have:

$$(\mathcal{M}, u) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]\varphi$$

(AC(b^u) requires this expression to hold for all subsets of $\vec{W}$)

The Need for Modification

Idea: Hold constant values $\vec{z}^*$ along the causal path. **AC2(b^o) 'Sufficiency'**: If $\vec{z}^*$ is such that $(\mathcal{M}, u) \models \vec{Z} = \vec{z}^*$, then for all subsets $\vec{Z}'$ of $\vec{Z} \setminus \vec{X}$, we have:

$$(\mathcal{M}, u) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]\varphi$$

(AC(b^u) requires this expression to hold for all subsets of $\vec{W}$)

The Issue: Due to the setting of $\vec{w}$ some variables in $\vec{Z}$ may change.
$\rightarrow$ In AC2(a), $\vec{w}$ could be *different* from the actual values observed in $(\mathcal{M}, u)$.

The Need for Modification

Idea: Hold constant values $\vec{z}^*$ along the causal path. **AC2(b^o) 'Sufficiency'**: If $\vec{z}^*$ is such that $(\mathcal{M}, u) \models \vec{Z} = \vec{z}^*$, then for all subsets $\vec{Z}'$ of $\vec{Z} \setminus \vec{X}$, we have:

$$(\mathcal{M}, u) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]\varphi$$

(AC(b^u) requires this expression to hold for all subsets of $\vec{W}$)

The Issue: Due to the setting of $\vec{w}$ some variables in $\vec{Z}$ may change.

$\rightarrow$ In AC2(a), $\vec{w}$ could be *different* from the actual values observed in $(\mathcal{M}, u)$.

$\rightarrow$ Allowing $\vec{W}$ to take non-actual values leads to 'fanciful' counterfactuals.

The Need for Modification

Idea: Hold constant values $\vec{z}^*$ along the causal path. **AC2(b^o) 'Sufficiency'**: If $\vec{z}^*$ is such that $(\mathcal{M}, u) \models \vec{Z} = \vec{z}^*$, then for all subsets $\vec{Z}'$ of $\vec{Z} \setminus \vec{X}$, we have:

$$(\mathcal{M}, u) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]\varphi$$

(AC(b^u) requires this expression to hold for all subsets of $\vec{W}$)

The Issue: Due to the setting of $\vec{w}$ some variables in $\vec{Z}$ may change.
→ In AC2(a), $\vec{w}$ could be *different* from the actual values observed in $(\mathcal{M}, u)$.
→ Allowing $\vec{W}$ to take non-actual values leads to 'fanciful' counterfactuals.

AC2(b^o) and AC2(b^u) can not really fix this. The 'issue' lies in the def. of AC2(a):
→ Directly modify AC2(a)→AC2(a^m)

The Modified HP Definition (mHP)

The *modified AC2* restricts the contingency $\vec{W}$:

AC2(a^m): There is a set of variables $\vec{W} \subseteq \mathcal{V}$ and a value x' such that if $(\mathcal{M}, u) \models \vec{W} = \vec{w}^*$, then:

$$(\mathcal{M}, u) \models [X \leftarrow x', \vec{W} \leftarrow \vec{w}^*]\neg\varphi$$

Key Restriction: The set of variables $\vec{W}$ (the "witnesses") **must** take the values $\vec{w}^*$, they took in the actual world $(\mathcal{M}, u)$. We no longer allow "off-path" interventions for the background variables.

The Modified HP Definition (mHP)

Modified HP Definition

A primitive event $\vec{X} = \vec{x}$ is an **actual cause** of φ in (M, u) if:

AC1 $(\mathcal{M}, u) \models (\vec{X} = \vec{x})$ and $(\mathcal{M}, u) \models \varphi$.

'The cause and the effect must actually happen.'

AC2 $(\mathcal{M}, u) \models [X \leftarrow x', \vec{W} \leftarrow \vec{w}^*]\neg\varphi$

x' must change φ, while $\vec{w}^*$ must take the values of the actual world.

AC3 $\vec{X}$ is minimal.

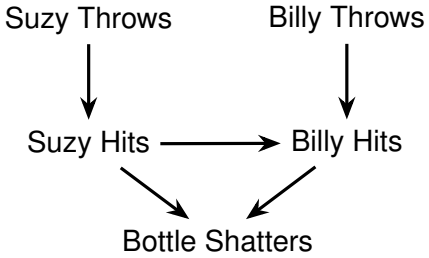
No subset of $\vec{X}$ satisfies AC1 and AC2.

Solving Rock-Throwing

Scenario:

Suzy and Billy throw ($ST = 1, BH = 1$).
$\vec{X} = \{ST\}$, $\varphi = (BS = 1)$.

1.) Set the witness $W = \{BH\}$. In the actual world, $BH = 0$ (because Suzy hit first $(\mathcal{M}, u) \models BH = 0$).



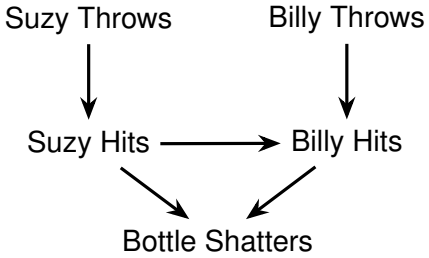
Solving Rock-Throwing

Scenario:

Suzy and Billy throw ($ST = 1, BH = 1$).
$\vec{X} = \{ST\}$, $\varphi = (BS = 1)$.

1.) Set the witness $W = \{BH\}$. In the actual world, $BH = 0$ (because Suzy hit first $(\mathcal{M}, u) \models BH = 0$).

2.) Under the contingency that BH is fixed to 0, if Suzy hadn't thrown ($ST = 0$), then BS would be 0.



Solving Rock-Throwing

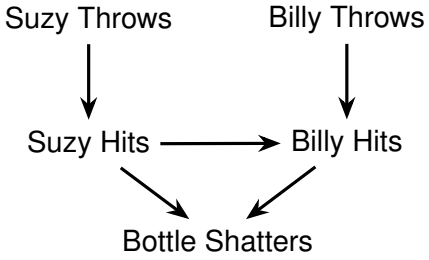
Scenario:

Suzy and Billy throw ($ST = 1, BH = 1$).
$\vec{X} = \{ST\}$, $\varphi = (BS = 1)$.

1.) Set the witness $W = \{BH\}$. In the actual world, $BH = 0$ (because Suzy hit first $(\mathcal{M}, u) \models BH = 0$).

2.) Under the contingency that BH is fixed to 0, if Suzy hadn't thrown ($ST = 0$), then BS would be 0.

3.1) **Billy is not a cause** because even if BH is fixed to its actual value (0), changing $BT = 0$ does nothing.



Solving Rock-Throwing

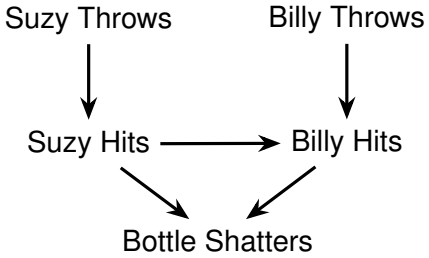
Scenario:
Suzy and Billy throw ($ST = 1, BH = 1$).
$\vec{X} = \{ST\}$, $\varphi = (BS = 1)$.

1.) Set the witness $W = \{BH\}$. In the actual world, $BH = 0$ (because Suzy hit first $(\mathcal{M}, u) \models BH = 0$).

2.) Under the contingency that BH is fixed to 0, if Suzy hadn't thrown ($ST = 0$), then BS would be 0.

3.1) **Billy is not a cause** because even if BH is fixed to its actual value (0), changing $BT = 0$ does nothing.

3.2) **Suzy *is* a cause** because under the actual evidence that $BH = 0$ changing $ST = 0$, entails $BS = 0$.



Normality

A *default* is the notion of what normally happens when no additional information is given.

$\rightarrow$ 'Birds typically fly.'

$\rightarrow$ 'A doctor typically treats an illness.'

Normality

A *default* is the notion of what normally happens when no additional information is given.

$\rightarrow$ 'Birds typically fly.'

$\rightarrow$ 'A doctor typically treats an illness.'

Some worlds $(\mathcal{M}, u)$ are more 'normal' than others.

$\rightarrow$ According to some particular preorder $\succeq$ of worlds.

$\rightarrow$ Extended Causal Models $\mathcal{M} = (\mathcal{S}, \mathcal{F}, \succeq)$

Normality

A *default* is the notion of what normally happens when no additional information is given.

$\rightarrow$ 'Birds typically fly.'

$\rightarrow$ 'A doctor typically treats an illness.'

Some worlds $(\mathcal{M}, u)$ are more 'normal' than others.

$\rightarrow$ According to some particular preorder $\succeq$ of worlds.

$\rightarrow$ Extended Causal Models $\mathcal{M} = (\mathcal{S}, \mathcal{F}, \succeq)$

Might be related to rate morality or guilt. …however, a car engine not working is not morally wrong, but fails 'to live up to a certain standard.'

Normality

A *default* is the notion of what normally happens when no additional information is given.

$\rightarrow$ 'Birds typically fly.'

$\rightarrow$ 'A doctor typically treats an illness.'

Some worlds $(\mathcal{M}, u)$ are more 'normal' than others.

$\rightarrow$ According to some particular preorder $\succeq$ of worlds.

$\rightarrow$ Extended Causal Models $\mathcal{M} = (\mathcal{S}, \mathcal{F}, \succeq)$

Might be related to rate morality or guilt. …however, a car engine not working is not morally wrong, but fails 'to live up to a certain standard.'

Worlds requiring lots of non-standard assumptions might be considered 'exotic'.

$\rightarrow$ Filter considered worlds by modifying $AC2(a)$:

$$AC2^{+}(a) := AC2(a^{m}) \text{ and } s_{\vec{X}=\vec{x},\vec{W}=\vec{W}^{*},\vec{u}} \succeq s_{\vec{u}}$$

"Require the intervened world to be at least as probable as the observed one."

Responsibility and Blame

Consider a vote of 11 people.

$\rightarrow$ Consider the vote succeeding with $6/5$.

$\rightarrow$ Consider the vote succeeding with $11/0$.

Responsibility and Blame

Consider a vote of 11 people.

$\rightarrow$ Consider the vote succeeding with $6/5$.

$\rightarrow$ Consider the vote succeeding with $11/0$.

So far actual causality was ‘all-or-nothing’.

$\rightarrow$ No consideration of probabilities.

$\rightarrow$ However, each voter might be less responsible for the vote succeeding in the $11/0$ scenario.

Responsibility and Blame

Consider a vote of 11 people.

$\rightarrow$ Consider the vote succeeding with $6/5$.

$\rightarrow$ Consider the vote succeeding with $11/0$.

So far actual causality was 'all-or-nothing'.

$\rightarrow$ No consideration of probabilities.

$\rightarrow$ However, each voter might be less responsible for the vote succeeding in the $11/0$ scenario.

(Naïve) **Responsibility** might be considered as the minimal set of changes/interventions that needs to be applied for the vote to change.

In particular, in a world where $\vec{X} = \vec{x}$ is a cause of φ under witnesses $\vec{W}$, $\vec{X} = \vec{x}$ has responsibility $dr^o((\mathcal{M}, \vec{u}), \vec{X} = \vec{x}, \varphi) := 1/(k+1)$ where $k = |\vec{W}|$.

"How many variables do I need to hold fixed for φ to happen."

Root Causes: A Probabilistic View

Root Cause Analysis consider the underlying factors for a particular observation that is considered unusual or out of distribution (o.o.d.).

Root Causes: A Probabilistic View

Root Cause Analysis consider the underlying factors for a particular observation that is considered unusual or out of distribution (o.o.d.).

An **Anomalous Event** is a realization of variables $\mathbf{V} = \mathbf{v}$ that deviates from the expected distribution $P(V)$.

A **Root Cause** is a (minimal) subset of exogenous noise variables $\mathbf{U}_{rc} \subseteq \mathbf{U}$ that is 'responsible' for the anomaly.

Root Causes: A Probabilistic View

Root Cause Analysis consider the underlying factors for a particular observation that is considered unusual or out of distribution (o.o.d.).

An **Anomalous Event** is a realization of variables $\mathbf{V} = \mathbf{v}$ that deviates from the expected distribution $P(V)$.

A **Root Cause** is a (minimal) subset of exogenous noise variables $\mathbf{U}_{rc} \subseteq \mathbf{U}$ that is 'responsible' for the anomaly.

Root Causes: In ‘standard’ SCM

Finding Root Causes $\mathbf{U}_{rc}$:

1.) A standard SCM $\mathcal{M}$ entails an expected distribution $P(\mathbf{V})$ over variables $\mathbf{V}$.
$\rightarrow$ For every observed vector $\mathbf{v}$, we can compute $P(\mathbf{v})$ according to $\mathcal{M}$.

Root Causes: In ‘standard’ SCM

Finding Root Causes $\mathbf{U}_{rc}$:

1.) A standard SCM $\mathcal{M}$ entails an expected distribution $P(\mathbf{V})$ over variables $\mathbf{V}$.
$\rightarrow$ For every observed vector $\mathbf{v}$, we can compute $P(\mathbf{v})$ according to $\mathcal{M}$.

2.) If for some $V \in \mathbf{V}$, $P(v)$ drops below a certain threshold we consider them out of distribution.

Question: Is a value o.o.d. due to anomalous parents or due to outlier noise u_i?

Root Causes: In ‘standard’ SCM

Finding Root Causes $\mathbf{U}_{rc}$:

1.) A standard SCM $\mathcal{M}$ entails an expected distribution $P(\mathbf{V})$ over variables $\mathbf{V}$.
$\rightarrow$ For every observed vector $\mathbf{v}$, we can compute $P(\mathbf{v})$ according to $\mathcal{M}$.

2.) If for some $V \in \mathbf{V}$, $P(v)$ drops below a certain threshold we consider them out of distribution.

3.) Structural equations induce a form of ‘normality’.
$\rightarrow$ Consider the conditional probability $p(v_i|pa(V_i))$.
$\rightarrow$ It is *high*, if the observed v_i follows the structural equation.
$\rightarrow$ It is *low*, if the observed v_i diverts from f_i due to noise.

Root Causes: In ‘standard’ SCM

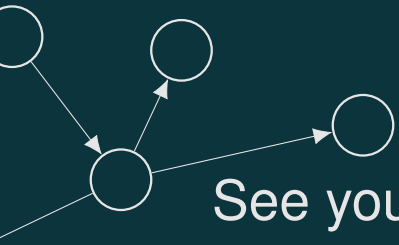
Finding Root Causes $\mathbf{U}_{rc}$:

1.) A standard SCM $\mathcal{M}$ entails an expected distribution $P(\mathbf{V})$ over variables $\mathbf{V}$.
$\rightarrow$ For every observed vector $\mathbf{v}$, we can compute $P(\mathbf{v})$ according to $\mathcal{M}$.

2.) If for some $V \in \mathbf{V}$, $P(v)$ drops below a certain threshold we consider them out of distribution.

3.) Structural equations induce a form of ‘normality’.
$\rightarrow$ Consider the conditional probability $p(v_i|pa(V_i))$.
$\rightarrow$ It is *high*, if the observed v_i follows the structural equation.
$\rightarrow$ It is *low*, if the observed v_i diverts from f_i due to noise.

- Marginal probabilities $P(v_i)$ let us detect anomalous variable values.
- Conditional probabilities $p(v_i|pa(V_i))$ tell us the locations of outlier noise.



See you next week!