Causes and Explanations:
A Structural-Model Approach

Joe Halpern
Cornell University

Judea Pearl
UCLA

Causal Reasoning

What does it mean that

- event A *causes* event B?
- event A is an *explanation* of B?

Giving precise and useful definitions of causality and explanation has proved notoriously difficult.

Why should we care?

- Solve longstanding philosophical problem.
- Critical in legal arguments:
  - Causality needed for determining responsibility.
- Automatic generation of adequate explanations is critical in planning, diagnosis, natural language understanding.

Outline

- A definition of actual causality in terms of *structural equations* (which uses *counterfactuals*)
  - Whether A causes B is relative to a model.
  - This moves the debate about causality to the right arena: do you have the right structural model?
- Showing that this definition handles well all the standard problematic examples in the literature.
- A definition of *explanation* relative to an epistemic state, using causality.
  - Roughly, A explains B if A causes B in all the situations an agent considers possible.

Causality: Intuition

[Lewis:] Basic intuition involves counterfactuals

- If A hadn't happened, B would not have happened

Typical (well-known problem): preemption

> [Hall] Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's throw had not preempted it.

So why is Suzy's throw the cause?

- If Suzy hadn't thrown under the contingency that Billy also didn't throw, then the bottle would not have shattered.

But then why isn't Billy's throw also a cause?

- Because it didn't hit the bottle.
- More generally, must restrict contingencies somehow.

Structural Equations

Idea: some random variables (propositions) have a causal effect on others.

- This effect is modeled by *structural equations*.

Split the random variables into

- *exogenous* variables, whose values are determined by factors outside the model
  - these values are taken as given
- *endogenous* variables.

The structural equations describe the values of the endogenous variables in terms of exogenous variables and other endogenous variables.

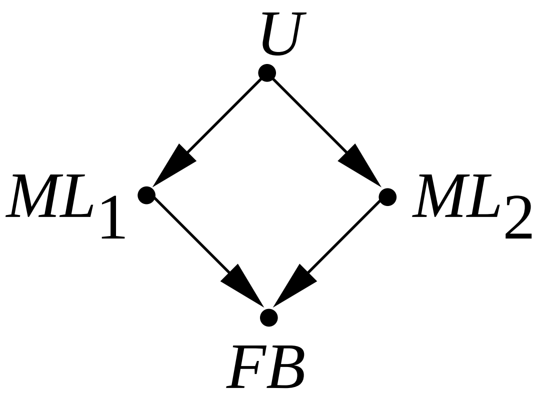
- One equation F_X for each endogenous variable X, describing X in terms of other variables.
  - $F_X(Y, U) = Y + U$
  - Usually write $X = Y + U$.
  - Does not mean $Y = X - U$!

Example 1: Arsonists

Two arsonists drop lit matches in different parts of a dry forest, and both cause trees to start burning. Consider two scenarios.

1. Disjunctive scenario: either match by itself suffices to burn down the whole forest.
2. Conjunctive scenario: both matches are necessary to burn down the forest

Both scenarios have the same causal network:



- endogenous variables ML_i, $i = 1, 2$:
  - $ML_i = 1$ iff arsonist i drops a match
- exogenous variable $U = (j_1 j_2)$ determines the motivation and state of mind of the arsonists:
  - $j_i = 1$ iff arsonist i intends to start a fire.
  - Actual context is $U = (11) \Rightarrow ML_1 = ML_2 = 1$
- endogenous variable FB (forest burns down).
- For the disjunctive scenario
  - $F_{FB}(u, 1, 1) = F_{FB}(u, 0, 1) = F_{FB}(u, 1, 0) = 1$
  - $F_{FB}(u, 0, 0) = 0$.
- For the conjunctive scenario
  - $F_{FB}(u, 1, 1) = 1$
  - $F_{FB}(u, 0, 0) = F_{FB}(u, 1, 0) = F_{FB}(u, 0, 1) = 0$.

Causal models

A *causal model* is a tuple $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$:

- $\mathcal{U}$: set of exogenous variables
- $\mathcal{V}$: set of endogenous variables
- $\mathcal{F}$: set of structural equations (one for each $X \in \mathcal{V}$):

$M_{\vec{X} \leftarrow \vec{x}} = (\mathcal{U}, \mathcal{V}, \mathcal{F}')$ is the causal model that results by deleting the equation for $\vec{X}$ and setting the variables in $\vec{X}$ to $\vec{x}$:

- Get F'_Y from F_Y by setting the values of the variables in $\vec{X}$ to $\vec{x}$.

(Some features of a) causal model M can be described by a *causal network*:

- Like Bayesian network, but edges interpreted causally

We restrict to causal models where all equations have a unique solution for each context $\vec{u}$:

- automatically holds in acyclic causal networks.

Reasoning about causality

We use the following language:

- primitive events $X = x$
- $[\vec{X} \leftarrow \vec{x}]\varphi$ ("after setting $\vec{X}$ to $\vec{x}$, φ holds")
  - This is essentially the same as the conditional implication $(\vec{X} = \vec{x}) \rightarrow \varphi$.
- close off under conjunction and negation.

Semantics:

- $M, \vec{u} \models Y = y$ if $Y = y$ in the unique solution to the equations in context $\vec{u}$
- $M, \vec{u} \models [\vec{X} \leftarrow \vec{x}]\varphi$ if $M_{\vec{X}=\vec{x}}, \vec{u} \models \varphi$.

Defining Causality

We want to define "A is the cause of B" (in context $\vec{u}$ of model M).

- We are assuming that all the relevant facts—the structural model and the context—are given.
- Which ones are the causes?

We restrict causes to conjunctions of primitive events:

$$X_1 = x_1 \wedge \ldots \wedge X_k = x_k$$

usually abbreviated as

$$\vec{X} = \vec{x}.$$

- One conjunct enough [Eiter-Lukasiewicz]
- Claim: OK to disallow disjunctive causes
- No need for probability, since everything given.

The events that can be caused are arbitrary Boolean combinations φ of primitive events.

(Preliminary) formal definition

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

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

- Both $\vec{X} = \vec{x}$ and φ are true in the actual world.

AC2. $\exists$ partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that if $(M, u) \models \vec{Z} = \vec{z}^*$, then

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

- changing $\vec{X}$ can change φ
- standard counterfactual clause, except we allow $\vec{W} = \vec{w}'$ [*structural contingency*]

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]\varphi$ for all subsets $\vec{Z}'$ of $\vec{Z}$.

- $\vec{Z}$ describes the *active causal process*.
- Setting $\vec{X}$ back to $\vec{x}$ forces φ to hold, even if $\vec{W} = \vec{w}'$ and some variables in the active causal process have their original values.

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

- No irrelevant conjuncts.

Arsonists Revisited

Each of $ML_1 = 1$ and $ML_2 = 1$ is a cause of $FB = 1$ in both scenarios.

To show that $ML_1 = 1$ is a cause in the disjunctive scenario: let $\vec{Z} = \{ML_1, FB\}$, so $\vec{W} = \{ML_2\}$.

- setting $ML_2 = 0$ satisfies AC2.
  - $ML_1 = 0 \Rightarrow FB = 0$; $ML_1 = 1 \Rightarrow FB = 1$.
- Need to use the structural contingency $ML_2 = 0$.
  - If $ML_2 = 1$, then $FB = 1$, independent of ML_1.
- Don't need structural contingency in the conjunctive scenario.

Note that ML_1 + sneezing is *not* a cause of $FB = 1$

- It passes AC1 and AC2, but not the minimality condition AC3.

Example 2: Preemption

[Hall:] Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's throw had not preempted it.

A naive causal model:

- ST for "Suzy throws", with values 0 (Suzy does not throw) and 1 (she does);
- BT for "Billy throws", with values 0 (he doesn't) and 1 (he does);
- BS for "bottle shatters", with values 0 (it doesn't shatter) and 1 (it does).

Problem: BT and ST play symmetric roles; there is nothing to distinguish them.

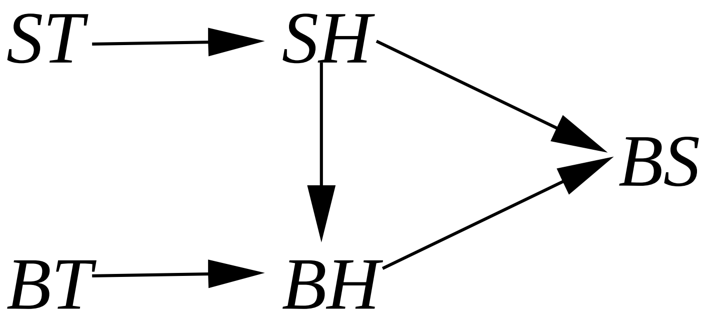
- Both $BT = 1$ and $ST = 1$ are causes in this model.

A Better Model

A more useful choice is to add two new random variables to the model:

- BH for "Billy's rock hits the (intact) bottle", with values 0 (it doesn't) and 1 (it does); and
- SH for "Suzy's rock hits the bottle", again with values 0 and 1.

Here is the causal network:



Now it is the case that $ST = 1$ is a cause of $BS = 1$, but $BT = 1$ is not (it fails AC2.)

- **Moral:** If there are redundant (potential) causes, we need a variable that distinguishes the two.

Example 3: Medical Treatment

[Hall:] Billy contracts a serious but nonfatal disease. He is treated on Monday, so is fine Tuesday morning. Had Monday's doctor forgotten to treat Billy, Tuesday's doctor would have treated him, and he would have been fine Wednesday morning. The catch: one dose of medication is harmless, but two doses are lethal.

Is the fact that Tuesday's doctor did *not* treat Billy the cause of him being alive (and recovered) on Wednesday morning?

The causal model has three random variables:

- MT (Monday treatment): 1–yes; 0–no
- TT (Tuesday treatment): 1–yes; 0–no
- BMC (Billy's medical condition):
  - 0–OK Tues. and Wed. morning,
  - 1–sick Tues. morning, OK Wed. morning,
  - 2–sick both Tues. and Wed. morning,
  - 3–OK Tues. morning, dead Wed. morning

Now have obvious equations.

Conclusions:

- $MT = 1$ is a cause of $BMC = 0$ and of $TT = 0$
- $TT = 0$ is a cause of Billy's being alive ($BMC = 0 \vee BMC = 1 \vee BMC = 2$).
- $MT = 1$ is *not* a cause of Billy's being alive (it fails condition AC2(a))

Conclusion: causality is *not* transitive nor does it satisfy right weakening.

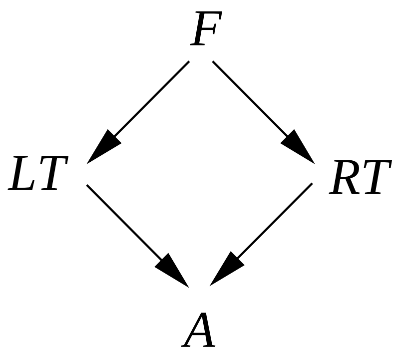
- Lewis assumes right weakening and forces transitivity.

Example 4: Railroad switches

[Hall:] A train comes to a fork in the railroad tracks. If you flip the switch, the train goes down the left-hand track; if you leave it where it is, the train goes down the right-hand track. Either way, the train will arrive at the same point, since the tracks reconverge.

Conclusion: flipping the switch is *not* a cause of the train's arrival in the obvious causal model:

However, it is a cause in a different causal model, with a different variable for each track:



This is a feature, not a bug!

- The change to a two-variable model represents a profound change to the story.

A More Refined Definition

Train example shows that adding new variables may change what counts as a cause.

- Refine notion of causality to be relative to a set of *allowed settings*.
  - In AC2, can only use allowed setting $(\vec{x}', \vec{w}')$.
  - In the train example with variables LT and RT, is $LT = RT = 0$ allowed?
- Restricting to only allowed settings handles other problems too; see paper.

Explanation

The definition of actual cause assumes that the causal model and all the relevant facts are given; the problem is to figure out which of the given facts are causes.

The role of explanation is to provide the information needed to establish causation.

- What counts as an explanation depends on an agent's *epistemic state* [Gärdenfors]
  - If the causal model is known, the epistemic state is a set of possible contexts.
  - In general, want to allow uncertainty about the causal model itself.
- Explanations (unlike causes) can be graded:
  - likelihood of explanation
  - how well explanation explains what is observed

Intuition: $\vec{X} = \vec{x}$ is an explanation of φ if relative to set $\mathcal{K}$ of contexts if:

- $\vec{X} = \vec{x}$ causes φ in each context in $\mathcal{K}$
- both $\vec{X} = \vec{x}$ and $\vec{X} \neq \vec{x}$ possible (relative to $\mathcal{K}$)

Explanation: The Formal Definition

Given a structural model M, $\vec{X} = \vec{x}$ *is an explanation of* φ *relative to a set* $\mathcal{K}$ *of contexts* if

EX1. $(M, \vec{u}) \models \varphi$ for each context $\vec{u} \in \mathcal{K}$.

- The agent *knows* φ
- (Known) observations require explanation

EX2. For all $\vec{u} \in \mathcal{K}$, if $(M, \vec{u}) \models \vec{X} = \vec{x}$, then $\vec{X} = \vec{x}$ is a *sufficient cause* of φ in $(M, \vec{u})$

- $\vec{X} = \vec{x}$ satisfies AC1 and AC2, but not necessarily minimality condition AC3

EX3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies EX2.

EX4. $(M, \vec{u}) \models \neg(\vec{X} = \vec{x})$ for some $\vec{u} \in \mathcal{K}$ and $(M, \vec{u}') \models \vec{X} = \vec{x}$ for some $\vec{u}' \in \mathcal{K}$.

- the explanation is neither already known nor vacuous.

The Arsonists Again

What counts as an explanation depends on $\mathcal{K}$. Consider the arsonist example, and suppose that arsonists are the only possible cause of fire:

- in the disjunctive scenario, both $ML_1 = 1$ and $ML_2 = 2$ are possible explanations (assuming that $\mathcal{K}$ includes the contexts where either or both arsonists started the fire).
- in the conjunctive scenario, there is no explanation (both arsonists must have lit a match), assuming that there are only two potential arsonists.
  - As soon as the fire is observed, the cause is known; can't satisfy EX4.
- If there are other possible causes of fire, then $ML_1 = 1 \wedge ML_2 = 1$ qualifies as an explanation in the conjunctive scenario.

Comparing Explanations

Not all explanations are considered equally good. Some explanations are more likely than others.

We can talk about the likelihood of an explanation by putting a probability Pr on $\mathcal{K}$, the set of possible contexts.

- The likelihood of the explanation $\vec{X} = \vec{x}$ is the probability of the set of contexts where $\vec{X} = \vec{x}$ is true.
- Compare $\Pr(ML_1 = 1)$ to $\Pr(ML_2 = 1)$ to see which is the better explanation in the disjunctive scenario.

Note:

- Let $\Pr^-$ be the pre-observation probability (e.g., before observing the forest fire):
  - $\Pr = \Pr^-(\cdot|\varphi)$ (i.e., Pr is the posterior).

  $\Pr(ML_1)$ really measures that prior probability of the explanation, given the observation $FB = 1$.

But using priors is not the only way to compare explanations ...

Explanatory Power

Suppose that we are considering a forest fire at a high altitude, with the disjunctive model.

- There is a small chance that there is insufficient oxygen for the fire to burn.
- If oxygen is an endogenous variable, it becomes a cause and an explanation of the forest fire,
- Moreover, $\Pr(O = 1)$ is high!
  - Higher than $\Pr(ML_1 = 1)$ or $\Pr(ML_2 = 1)$.

But oxygen does not seem like a very good explanation ...

Formal definition:

- Let $\mathcal{K}_{\vec{X}=\vec{x},\varphi} = \{\text{contexts: } \vec{X} = \vec{x} \text{ causes } \varphi\}$.
- The *explanatory power* of $\vec{X} = \vec{x}$ is

$$\Pr{}^{-}(\mathcal{K}_{\vec{X}=\vec{x},\varphi} | \vec{X} = \vec{x}).$$

$O = 1$ has lower explanatory power than $ML_1 = 1$:

$$\Pr{}^{-}(\mathcal{K}_{O=1,FB=1} | O = 1) < \Pr{}^{-}(\mathcal{K}_{ML_1=1,FB=1} | ML_1 = 1)$$

There is a tension between the explanatory power of a partial explanation and its prior probability.

- The prior probability that $\vec{X} = \vec{x}$ vs. the probability that $\vec{X} = \vec{x}$ explains φ.
- Adding conjuncts increases the explanatory power and decreases the prior probability.

Explanation: A General Definition

In general, there may be uncertainty about the causal model:

- An epistemic state $\mathcal{K}$ consists not only of contexts, but of *situations* $(M, \vec{u})$:
  - model M + context $\vec{u}$.

An explanation should include causal information, such as "prayers do not cause fires".

A *(general) explanation* has the form $(\psi, \vec{X} = \vec{x})$, where ψ is an arbitrary causal formula.

- Intuition: $\vec{X} = \vec{x}$ is a cause of φ in all causal models satisfying ψ.
  - See paper for the full general definition (very similar in spirit to the basic definition)

Related Work

Causality:

- *Lots* of work in the philosophy literature
- For causality, most prominent is Lewis's work
  - Lots of follow-on work
  - Our definition seems to give more reasonable answers than Lewis's in the standard problematic examples.
- Pearl's *causal beam*
  - This definition more transparent, solves a number of problems in that definition.
- Tian and Pearl
  - Consider *necessary causes*, not actual causes
    - B would not have occurred if A had not occurred
- Heckerman and Shachter: consider when one random variable causes another
- formal action theory [Lin; Sandewall; ...]—different focus: incorporating causal relationships into a knowledge base.

Explanation:

- Again, *lots* of work in philosophy literature.
- Main names are Hempel, Salmon, Lewis
- Gärdenfors stressed importance of explanation relative to epistemic state
- Philosophers seem reluctant to build definition of explanation on top of definition of causality.
  - Result: no notion of "directionality"
  - Does the length of the shadow explain the height of the flagpole?
- Dominant approach in AI literature: *maximum a posteriori approach* [Pearl; Shimony; Druzdel/Henrion]
  - focuses on prior, but does not take explanatory power into account (see [Chajewska/Halpern])

Summary

The counterfactual approach to causation and explanation need not be abandoned, but ...

- Must deal with contingencies carefully.
- Need to be careful about the choice of model.
- Explanation is relative to knowledge.

Problems:

- Dealing with disjunctive explanations
- Resolving the tension between explanatory power and goodness of explanation