Actual Causality: A Survey

Joe Halpern
Cornell University

Includes joint work with Judea Pearl (UCLA), Hana Chockler (King's College), and Chris Hitchcock (Cal Tech)

The big picture

Defining causality is notoriously difficult.

- Many approaches have been considered in the philosophy and legal literatures for both
  - *type causality*: smoking causes cancer
  - *token/actual causality*: the fact that Willard smoked for 30 years caused him to get cancer

The big picture

Defining causality is notoriously difficult.

- Many approaches have been considered in the philosophy and legal literatures for both
  - *type causality*: smoking causes cancer
  - *token/actual causality*: the fact that Willard smoked for 30 years caused him to get cancer

Why should we care?

> *It's true that it was pouring rain last night, and I was drunk, but the cause of the accident was the faulty brakes in the car (so I'm suing GM).*

The big picture

Defining causality is notoriously difficult.

- Many approaches have been considered in the philosophy and legal literatures for both
  - *type causality*: smoking causes cancer
  - *token/actual causality*: the fact that Willard smoked for 30 years caused him to get cancer

Why should we care?

> *It's true that it was pouring rain last night, and I was drunk, but the cause of the accident was the faulty brakes in the car (so I'm suing GM).*

- Issues of actual causality are omnipresent in the law.
- Historians and scientists are interested in causality
- Statisticians are very concerned with token causality.
- Causality is also relevant to computer science (!)

The big picture (cont'd)

What does it mean for A to be a cause of B?

- Attempts to define causality go back to Aristotle
- The modern view arguably dates back to Hume (1748)
- Relatively recent trend (going back to Lewis (1973)) to capturing actual causality: use counterfactuals
  - A is a cause of B if it is the case that if A had not happened, B would not have happened
  - If the brakes hadn't been faulty, I wouldn't have had the accident
- More recent trend: capture the counterfactuals using structural equations (Pearl 2000)
- Pearl and I gave a definition of actual causality using structural equations:
  - original definition: Halpern-Pearl, UAI 2001
  - improved (i.e., corrected): Halpern-Pearl, 2005 (BJPS)
  - yet another definition: Halpern, 2015 (IJCAI)

Why it's hard

The simple counterfactual definition doesn't always work

- When it does, we have what's called a *but-for cause*
- This is the situation considered most often in the law

Typical (well-known problem): preemption

> *[Lewis:] 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?

Why it's hard

The simple counterfactual definition doesn't always work

- When it does, we have what's called a *but-for cause*
- This is the situation considered most often in the law

Typical (well-known problem): preemption

> *[Lewis:] 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?

- Given that Billy doesn't hit the bottle, if Suzy hadn't thrown, then the bottle would not have shattered.

Why it's hard

The simple counterfactual definition doesn't always work

- When it does, we have what's called a *but-for cause*
- This is the situation considered most often in the law

Typical (well-known problem): preemption

> *[Lewis:] 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?

- Given that Billy doesn't hit the bottle, if Suzy hadn't thrown, 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! (Duh . . .)

But how do we capture this?

Structural equations

Idea: World described by variables that affect each other

- This effect is modeled by *structural equations*.

Split the random variables into

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

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

- Have an equation for each variable
  - $X = Y + U$ does not mean $Y = X - U$!

Reasoning about causality

Syntax: We use the following language:

- primitive events $X = x$
- $[\vec{X} \leftarrow \vec{x}]\varphi$ ("after setting $\vec{X}$ to $\vec{x}$, φ holds")
- close off under conjunction and negation.

Semantics: 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 $X \in \mathcal{V}$):
  - E.g., $X = Y \wedge Z$

Let $\vec{u}$ be a *context*: a setting of the exogenous variables:

- $(M, \vec{u}) \models Y = y$ if $Y = y$ is unique solution to equations in $\vec{u}$
- $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}]\varphi$ if $(M_{\vec{X} \leftarrow \vec{x}}, \vec{u}) \models \varphi$.
- $M_{\vec{X} \leftarrow \vec{x}}$ is the causal model after setting $\vec{X}$ to $\vec{x}$:
  - replace the original equations for the variables in $\vec{X}$ by $\vec{X} = \vec{x}$.

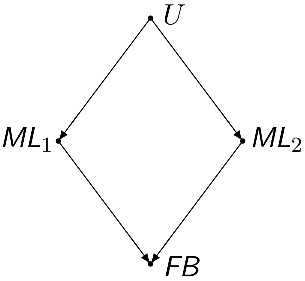
Example 1: Arsonists

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

Arsonist scenarios

Same causal network for both scenarios:



- endogenous variables ML_i, $i = 1, 2$:
  - $ML_i = 1$ iff arsonist i drops a match
- exogenous variable $U = (j_1 j_2)$
  - $j_i = 1$ iff arsonist i the background conditions are such that arsonist i will drop a match
- endogenous variable FB (forest burns down).
  - For the disjunctive scenario $FB = ML_1 \vee ML_2$
  - For the conjunctive scenario $FB = ML_1 \wedge ML_2$

Defining causality

We want to define "A is a cause of B" given $(M, \vec{u})$.

- Assuming all relevant facts—structural model and context—given.
- Which events 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 $\vec{X} = \vec{x}$.

- The conjunction is sometimes better thought of as a *disjunction*
  - This will be clearer with examples
- No need for probability, since everything given.

Arbitrary Boolean combinations φ of primitive events can be caused.

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. A somewhat complicated condition, capturing the counterfactual requirements.

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

- No irrelevant conjuncts.
- Don't want "dropping match and sneezing" to be a cause of the forest fire if just "dropping match" is.

AC2

In the original definition, AC2 was quite complicated. Now it's much simpler:

AC2. There is a set $\vec{W}$ of variables in $\mathcal{V}$ and a setting $\vec{x}'$ of the variables in $\vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w}$, then

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

In words: if we keep the variables in $\vec{W}$ fixed at their actual values, then changing $\vec{X}$ can change the outcome φ.

- So the counterfactual holds (if $\vec{X}$ weren't $\vec{x}$, then φ would not hold) provided the variables in $\vec{W}$ are held fixed to their actual values.

Example 1: Arsonists revisited

Each of $ML_1 = 1$ and $ML_2 = 1$ is a (but-for) cause of $FB = 1$ in the conjunctive scenario.

- If either arsonist hadn't dropped a match, there wouldn't have been a fire.
- An effect can have more than one cause.

Example 1: Arsonists revisited

Each of $ML_1 = 1$ and $ML_2 = 1$ is a (but-for) cause of $FB = 1$ in the conjunctive scenario.

- If either arsonist hadn't dropped a match, there wouldn't have been a fire.
- An effect can have more than one cause.

In the disjunctive scenario, $ML_1 = 1 \wedge ML_2 = 1$ is a cause:

- If we change both ML_1 and ML_2, the outcome changes.
- $ML_1 = 1$ is not a cause:
  - if we keep ML_2 fixed at its actual value, then no change in ML_1 can change the outcome; similarly for ML_1
- Similarly, $ML_2 = 1$ is not a cause

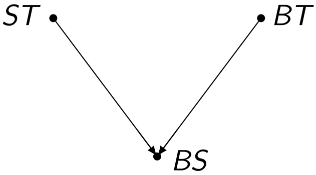
This seems inconsistent with natural language usage!

- Two ways to think about this:
  - What we typically call a cause in natural language is a conjunct of a cause according to this definition.
  - We can think of the disjunction $ML_1 = 1 \vee ML_2 = 1$ as a but-for cause of $FB = 1$

Example 2: Throwing rocks

A naive causal model looks just like the arsonist model:

- ST for "Suzy throws" (either 0 or 1)
- BT for "Billy throws" (either 0 or 1)
- BS for "bottle shatters" (either 0 or 1)



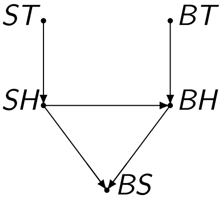
Problem: BT and ST play symmetric roles; nothing distinguishes them.

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

A better model

If we want Suzy to be a cause of the bottle shattering, and not Billy, there must be variables that distinguish Suzy from Billy:

- *SH* for "Suzy's rock hits the (intact) bottle"; and
- *BH* for "Billy's rock hits the (intact) bottle".
  - If Suzy hits ($\mathit{SH} = 1$) then Billy doesn't hit



Suzy is a cause because

$$(M, u) \models [\mathit{ST} = 0, \mathit{BH} = 0](\mathit{BS} = 0).$$

Billy is not a cause:

- There is nothing we can hold fixed at its actual value to make *BS* counterfactually depend on *BT*.

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

The equations are obvious.
What can we say about causality?

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

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

- Lewis assumes right weakening and forces transitivity.

Example 4: Normality

> *[Knobe and Fraser:] The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own. In practice, both assistants and faculty members take the pens. On Monday morning, both an assistant and Prof. Smith take pens. Later, the receptionist needs to take an important message, but there are no pens left on her desk.*

Who is the cause?

- Most people say Prof. Smith
- In the obvious causal model, both Prof. Smith and the assistant play completely symmetric roles.
  - They are both (but-for) causes.

Defaults and normality

There must be more to causality than just the structural equations.

Key insight

> *[Kahneman/Miller, 1986]: "an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it."*

We can formalize this by ordering worlds in terms of their "normality"

- This is a standard way of modeling default reasoning in AI
- The world where the Prof. Smith does not take the pen and the assistant does is more normal than the world where Prof. Smith takes it and the assistant doesn't.
  - Thus, we prefer to modify the world in the former way.
    - Prof. Smith is a "better" cause

A more refined definition [Halpern/Hitchcock]

Call $(\vec{X} = \vec{x}', \vec{W} = \vec{w})$ a *witness* for $\vec{X} = \vec{x}$ being a cause of φ in $(M, \vec{u})$ if $(M, w) \models \vec{W} = \vec{w}$ and

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

- $\vec{X} = \vec{x}'$, $\vec{W} = \vec{w}$ is a witness for AC2 holding.

There may be several witnesses for $\vec{X} = \vec{x}$ being a cause of φ. $\vec{X}_1 = \vec{x}_1$ is a *better* cause of φ than $\vec{X}_2 = \vec{x}_2$ if the the most normal witness for $\vec{X}_1 = \vec{x}_1$ being a cause of φ is more normal than the most normal witness for $\vec{X}_2 = \vec{x}_2$ being a cause of φ.

- We thus get a *graded* notion of causality
  - This can be used to capture a lot of human causal judgments
    - E.g., attenuation of responsibility along a causal chain

What is normality?

What does the normality ordering on witnesses represent:

- statistical frequency
- typicality
- conformance to a norm (as in the Knobe-Fraser example)
- moral obligations
- . . .

People seem to consider all these notions when judging normality.

- We take the normality ordering to be subjective
- Something that lawyers can argue over.

Responsibility and blame [Halpern/Chockler]

The definition of causality can be extended to deal with responsibility and blame (and explanation).

Causality is a 0-1 notion: either A causes B or it doesn't

- Can easily extend to talking about the *probability* that A causes B
  - Put a probability on contexts

But not all causes are equal:

- Suppose B wins an election against G by a vote of 11–0.
- Each voter for B is (part of) a cause of B's winning.
- However, it seems that their degree of responsibility should not be the same as in the case that the vote is 6–5.

Voting example

There are 11 voters and an outcome, so 12 random variables:

- $V_i = 0/1$ if voter i voted for G/B, for $i = 1, \dots, 11$;
- $O = 1$ if B has a majority, otherwise 0.

$V_1 = 1$ is a cause of $O = 1$ in a context where everyone votes for B.

- If $V_1, V_2, \dots, V_6$ are set to 0, then AC2 holds.

$V_1 = 1$ is also a cause of $O = 1$ in a context where only $V_1, \dots, V_6$ vote for B, so the vote is 6–5.

- Now only have to change the value of V_1 in AC2

Key idea: use the size of the smallest witness as a measure of degree of responsibility.

Responsibility: formal definition

The *degree of responsibility of* $X = x$ *for* φ *in* $(M, \vec{u})$ is

- 0 if $X = x$ is not part of a cause $\vec{X} = \vec{x}$ of φ in $(M, \vec{u})$;
- $1/k$ if $X = x$ is part of a cause of $\vec{X} = \vec{x}$ of φ in $(M, \vec{u})$ with a witness $(\vec{X} = \vec{x}', \vec{W} = \vec{w})$, such that $|\vec{X}| + |\vec{W}| = k$, and k is minimal
  - $X = x$ is not part of a cause $\vec{X}_1 = \vec{x}_1$ with witness $(\vec{X}_1 = \vec{x}'_1, \vec{W}_1 = \vec{w}_1)$ and $|\vec{X}_1| + |\vec{W}_1| < k$.

Example:

- If vote is 11–0, V_1 has degree of responsibility $1/6$
- If vote is 6–5, V_1 has degree of responsibility 1

Degree of blame

When determining responsibility, it is assumed that everything relevant about the facts of the world and how the world works is known.

- In the voting example, the vote is assumed known; no uncertainty.
- Also true for causality.

Sometime we want to take an agent's epistemic state into account:

- A doctor's use of a drug to treat a patient may have been the cause of a patient's death
- The doctor then has degree of responsibility 1.
- But what if he had no idea there would be adverse side effects?
  - He may then not be to blame for the death

In legal reasoning, what matters is not only what he did know, but what he *should have known*

We define a notion of degree of blame relative to an epistemic state

- The epistemic state is a set of situations
  - the situations the agents considers possible + a probability distribution on them
- Roughly speaking, the degree of blame is the expected degree of responsibility, taken over the situations the agent considers possible.

Blame: Example

Consider a firing squad with 10 excellent marksmen.

- Only one of them has live bullets in his rifle; the rest have blanks.
- The marksmen do not know which of them has the live bullets.
- The marksmen shoot at the prisoner and he dies.

Then

- Only marksman with the live bullets is the cause of death.
- That marksman has degree of responsibility 1 for the death.
- The others have degree of responsibility 0.
- Each marksmen has degree of blame $1/10$
  - This is the expected degree of responsibility.

Some CS applications

Causality is omnipresent in legal reasoning and scientific reasoning.

- The law does not have a good definition beyond the but-for definition, and they realize that they need it.

But there are also lots of other potential CS applications:

- *Accountability* [Datta et al; Feigenbaum, Jaggard, Wright]
    - What is the cause of the security breach?
- *Databases* [Gatterbauer, Meliou, Moore, Suciu]
    - interested in causality queries
    - What tuple *caused* a particular output to a query?
    - What caused the network failure?
- *Program verification* [Chockler,Halpern,Kupferman; Beer, Chockler, et al.; Chockler, Grumberg, Yadgar]
    - What is the cause of the counterexample to the spec?
    - Which line of code is most responsible?

Complexity

Unlike the typical examples in philosophy, the CS applications tend to involve large examples:

- hundreds of variables

How hard is to compute if A is a cause of B?
Theorem: The complexity of determining whether $\vec{X} = \vec{x}$ is a cause of φ in $(M, \vec{u})$ is D_1^P-complete.

- D_1^p [Papadimitriou-Yannakakis] consists of those language L such that $L = L_1 \cap L_2$, where L_1 is in NP, L_2 is in co-NP.
- Checking if AC2 holds is in NP; checking AC3 is in co-NP.

Theorem: The complexity of determining whether $X = x$ is a cause of φ in $(M, \vec{u})$ is NP-complete for binary models (i.e., models where all variables are binary).

The trouble with people

- People don't always agree on ascriptions of causality
- People apply multiple intuitions to inferring causality
  - looking for an "active" physical process
  - counterfactual reasoning

The trouble with people

- People don't always agree on ascriptions of causality
- People apply multiple intuitions to inferring causality
  - looking for an "active" physical process
  - counterfactual reasoning

Nevertheless, experiments on Amazon Turk with voting scenarios [Gerstenfeld/Halpern/Tenenbaum] show that

- the naive responsibility definition does predict qualitatively how people acribe responsibility in many situations
  - people's responsibility ascriptions are affected by normality considerations
  - and also affected by prior probabilities
    - People conflate responsibility and blame [Howe/Sloman],

Discussion

Depending on their focus, people give different answers.

- What is the cause of the traffic accident?
  - The engineer's answer: the bad road design
  - The mechanic's answer: bad brakes
  - The sociologist's answer: the pub near the highway
  - The psychologist's answer: the driver was depressed

  These answers are all reasonable.

  - It depends what we view as exogenous and endogenous

Discussion

Depending on their focus, people give different answers.

- What is the cause of the traffic accident?
  - The engineer's answer: the bad road design
  - The mechanic's answer: bad brakes
  - The sociologist's answer: the pub near the highway
  - The psychologist's answer: the driver was depressed

  These answers are all reasonable.
  - It depends what we view as exogenous and endogenous

Nevertheless, I am optimistic that we can find *useful* definitions.

- Taking normality into account seems to help.
- The structural models framework seems to be widely applicable.
  - It's been used to define blame, responsibility, explanation, blameworthiness, intention, and harm (so far . . .)
- I expect applications to drive much of the research agenda