



# Cause, Responsibility, and Blame

## A Structural-Model Approach

Joe Halpern

Cornell University

Joint work with Judea Pearl (causality);

Hana Chockler and Orna Kupferman  
(responsibility and blame)



# Outline

- A definition of actual causality in terms of *structural equations* (which uses *counterfactuals*) [H & Pearl]
  - 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 many standard problematic examples in the literature.
- Extending approach to responsibility and blame [Chocker & H]
- Applications to program testing [Chockler, H, & Kupferman]

# 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 have shattered.



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 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:** World described by random variables that affect each other

- This effect is modeled by *structural equations*.

# Structural Equations

**Idea:** World described by random 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

**Idea:** World described by random 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  $Y = U - X$ !



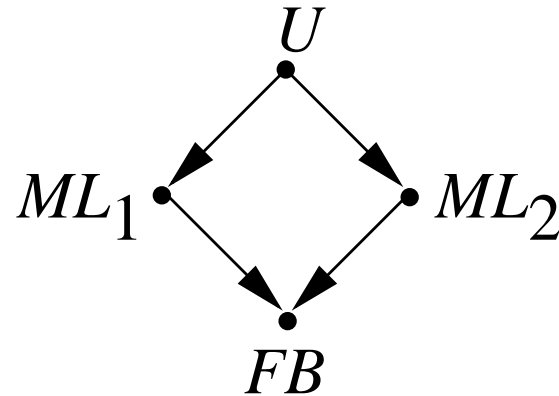
# Example 1: Arsonists

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  $i$  intends to start a fire.
- endogenous variable  $FB$  (forest burns down).
  - For the disjunctive scenario  $FB = ML_1 \vee ML_2$
  - For the conjunctive scenario  $FB = ML_1 \wedge ML_2$

# 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  $X \in \mathcal{V}$ ):

# 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  $X \in \mathcal{V}$ ):

(Some features of a) causal model 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

**Syntax:** We use the following language:

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

# Reasoning about causality

**Syntax:** We use the following language:

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

**Semantics:**

- $M, \vec{u} \models Y = y$  if  $Y = y$  in unique solution to equations in  $\vec{u}$
- $M, \vec{u} \models [\vec{X} \leftarrow \vec{x}]\varphi$  if  $M_{\vec{X}=\vec{x}}, \vec{u} \models \varphi$ .
- $M_{\vec{X} \leftarrow \vec{x}} = (\mathcal{U}, \mathcal{V} - \vec{X}, \mathcal{F}')$  is the causal model that results from
  - deleting the equations for variables  $\vec{X}$  and
  - getting new equation for  $Y \notin \vec{X}$  by setting variables in  $\vec{X}$  to  $\vec{x}$

# Defining Causality

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

- Assuming all relevant facts—structural model and context—given.
- Which events are the causes?

We restrict causes to conjunctions of primitive events:

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

usually abbreviated as  $\vec{X} = \vec{x}$ .

- One conjunct enough [Eiter-Lukasiewicz]
- No need for probability, since everything given.

Arbitrary Boolean combinations  $\varphi$  of primitive events can be caused.

# (Preliminary) formal definition

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

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

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



# (Preliminary) formal definition

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  $\varphi$

- standard counterfactual clause, except we allow  $\vec{W} = \vec{w}'$

[*structural contingency*]

# (Preliminary) formal definition

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  $\varphi$

- 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  $\vec{Z}' \subseteq \vec{Z}$ .

- $\vec{Z}$  describes the *active causal process*.

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

# (Preliminary) formal definition

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

# 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.

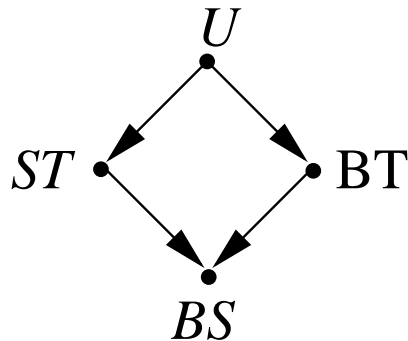
# Example 2: Preemption

[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.

# 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 looks just like the arsonist model:

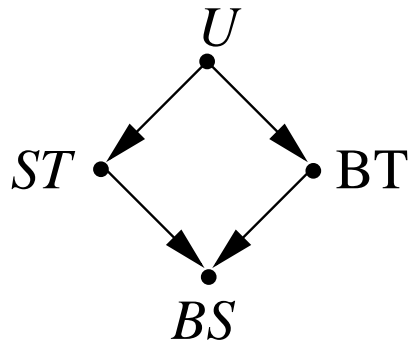


- *ST* for “Suzy throws” (either 0 or 1)
- *BT* for “Billy throws” (either 0 or 1)
- *BS* for “bottle shatters” (either 0 or 1)

# 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 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.

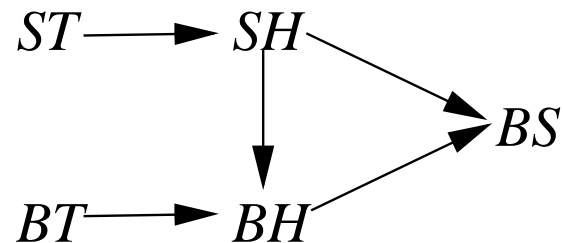
- Both  $BT = 1$  and  $ST = 1$  are causes in this model.

# A Better Model

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  $ST = 1$  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

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

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 ( $BMC = 0 \vee BMC = 1 \vee BMC = 2$ ).
- $MT = 1$  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.

# Degree of Responsibility

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 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 number of variables whose value has to change in AC2 as a measure of degree of responsibility.

# Responsibility: Formal Definition

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

- 0 if  $X = x$  is not a cause of  $\varphi$  in  $(M, \vec{u})$ ;
- $1/(k + 1)$  if  $X = x$  is a cause of  $\varphi$  in  $(M, \vec{u})$  and there exists a partition  $(\vec{Z}, \vec{W})$  and setting  $(x', \vec{w}')$  for which AC2 holds and
  - (1)  $k$  variables in  $\vec{W}$  have different values in  $\vec{w}'$  than they do in the context  $\vec{u}$
  - (2) We can't do better than  $k$ 
    - there is no partition  $(\vec{Z}', \vec{W}')$  and setting  $(x'', \vec{w}'')$  satisfying AC2 such that only  $k' < k$  variables have different values in  $\vec{w}''$  than they do the context  $\vec{u}$ .

# Responsibility: Formal Definition

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

- 0 if  $X = x$  is not a cause of  $\varphi$  in  $(M, \vec{u})$ ;
- $1/(k + 1)$  if  $X = x$  is a cause of  $\varphi$  in  $(M, \vec{u})$  and there exists a partition  $(\vec{Z}, \vec{W})$  and setting  $(x', \vec{w}')$  for which AC2 holds and
  - (1)  $k$  variables in  $\vec{W}$  have different values in  $\vec{w}'$  than they do in the context  $\vec{u}$
  - (2) We can't do better than  $k$

## 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

# 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.

# 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.

# 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.

# 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.

# Application: Coverage

*Model checking* tells you if a program satisfies a specification.

- If algorithm says no, it provides a counterexample.
- If algorithm says yes, then it terminates
  - Problem: what if there's an error in the spec?
- Recent emphasis on various sanity checks
  - *Coverage estimation*: which parts of the program are actually relevant for the spec.
  - An “unused” part of the program may signal an error.



# Coverage and Causality

Key observation: coverage is like causality

- Which parts of the program *cause* the spec to be satisfied?
- We can also measure the degree of responsibility of a node in a circuit for satisfying a spec
  - A low degree of responsibility might indicate a problem—part of the circuit is not so important
  - A high degree of responsibility says the node is critical—this could be a problem for fault tolerance

Groce et al. [2006] define a notion of *error explanation* also based on counterfactuals: is a line of code a cause for the error?

- Also can be extended by responsibility

# Conclusion

- The structural models approach can capture a number of intuitions rather naturally.
- The approach can be extended to deal with
  - explanation
  - degree of responsibility
  - blame
- These notions can be applied to verification.

There's much more that can be done!