Explaining actual causation in terms of possible causal processes

Bart Bogaerts (with Marc Denecker and Joost Vennekens)

JELIA'19

May 10, 2019

Type causality
“Smoking causes cancer”

Actual causality
“The cars faulty brakes caused the crash”

Formalizing actual causation

What does it mean to say:

C caused E?

Content

1. Actual Causality: Structural Equations and Counterfactual Reasoning
2. Our Approach to Actual Causality
3. Conclusion

The counterfactual approach

- Lewis (1973):

  C caused E := “Without C, E would not have been”

- Pearl (2000)
- Halpern & Pearl (2001, 2005) (HP)
- Halpern (2016), Fenton-Glynn (2015), Gerstenberg (2015), Vennekens (2011), …
- Counterfactual definitions of actual causation in the context of structural equation models.

Example

Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both are expert rock-throwers, Billy's would have shattered the bottle had it not been preempted by Suzy's throw.

$$\begin{aligned} \mathit{Shatter} &:= \mathit{SuzyHits} \vee \mathit{BillyHits} \\ \mathit{SuzyHits} &:= \mathit{SuzyThrows} \\ \mathit{BillyHits} &:= \mathit{BillyThrows} \wedge \neg \mathit{SuzyThrows} \end{aligned}$$

World: $\mathit{SuzyThrows} = \mathit{BillyThrows} = \mathbf{t}$. What are the actual causes of Shatter?

Definition (Halpern, 2016)

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

- $\bar{X} = \bar{x}$ and ϕ both hold in the world $(M, \bar{u})$
- There is a set of variables W such that if we fix their value and change X's value, ϕ no longer holds
- $\bar{X}$ is a minimal such set

Example

$$\begin{aligned} \mathit{Shatter} &:= \mathit{SuzyHits} \lor \mathit{BillyHits} \\ \mathit{SuzyHits} &:= \mathit{SuzyThrows} \\ \mathit{BillyHits} &:= \mathit{BillyThrows} \land \neg \mathit{SuzyThrows} \end{aligned}$$

In world $\mathit{SuzyThrows} = \mathit{BillyThrows} = \mathbf{t}$, *SuzyThrows* is an actual cause of *Shatter*:

- Contingency set $W = \{\mathit{BillyHits}\}$
- If Suzy does not throw under this intervention, the bottle does not shatter

Criticisms against counterfactual definitions

- Objections against counterfactual approach

  Alternative definitions

  Hall (2004), Baumgartner (2013), Bochman & Lifschitz (2015)

- Problematic causal scenarios for all counterfactual definitions
  $\Rightarrow$ Refinements of the definitions

Criticisms against counterfactual definitions

Halpern (2016b) analyzes 6 of these problematic causal scenarios.
$\rightarrow$ different informal interpretation of the same formal model
$\rightarrow$ HP correct under that interpretation!

Criticisms against counterfactual definitions

Halpern (2016b) analyzes 6 of these problematic causal scenarios.
$\rightarrow$ different informal interpretation of the same formal model
$\rightarrow$ HP correct under that interpretation!
Thus... causal models are ambiguous! Some knowledge is missing.

Ambiguities (Example 1, Scenario 1)

- Two deadly potions (Arsenic, Strychnine)
- They work independently

$$\mathit{Dead} := \mathit{Strychnine} \vee \mathit{Arsenic}$$

Ambiguities (Example 1, Scenario 2)

- Two deadly potions (Arsenic, Strychnine)
- But... Arsenic preempts the chemical process by which Strychnine poisoning works

$$Dead := (\neg Arsenic \wedge Strychnine) \vee Arsenic$$

Equivalent to:

$$Dead := Strychnine \vee Arsenic$$

Resolving the Ambiguitiy (Halpern)

- KR methodology: reify mechanisms by auxiliary variables

$$\mathit{Dead} := \mathit{SPoising} \vee \mathit{Arsenic}$$
$$\mathit{SPoisining} := \neg\mathit{Arsenic} \wedge \mathit{Strychnine}$$

- Works: gets the right answers, but...
- no principled explanation of actual causation in terms of the causal process and mechanisms

Ambiguities (Example 2)

An assassin may kill a victim by administering deadly poison. A bodyguard may rescue the victim by administering an antidote.

$$\mathit{Dead} := \mathit{Poison} \land \mathit{NoAntidote}$$

Ambiguities (Example 3, Scenario 1)

Halpern (2016b)

> *"There are four endogenous binary variables, A, B, C, and S, taking values 1 (on) and 0 (off). Intuitively, A and B are supposed to be alternative causes of C, and S acts as a switch. If S = 0, the causal route from A to C is active and that from B to C is dead; and if S = 1, the causal route from A to C is dead and the one from B to C is active."*

Ambiguities (Example 3, Scenario 1)

Halpern (2016b)

> *"There are four endogenous binary variables, A, B, C, and S, taking values 1 (on) and 0 (off). Intuitively, A and B are supposed to be alternative causes of C, and S acts as a switch. If S = 0, the causal route from A to C is active and that from B to C is dead; and if S = 1, the causal route from A to C is dead and the one from B to C is active."*

$$C := (\neg S \wedge A) \vee (S \wedge B)$$

Ambiguities (Example 3, Scenario 1)

Halpern (2016b)

> *"There are four endogenous binary variables, A, B, C, and S, taking values 1 (on) and 0 (off). Intuitively, A and B are supposed to be alternative causes of C, and S acts as a switch. If S = 0, the causal route from A to C is active and that from B to C is dead; and if S = 1, the causal route from A to C is dead and the one from B to C is active."*

$$C := (\neg S \wedge A) \vee (S \wedge B)$$

What is the actual cause of C, intuitively?

- when S, then A
- when $\neg S$, then B

Ambiguities (Example 3, Scenario 2)

Halpern (2016b)

> *"But now consider a slightly different story. This time, we view B as the switch, rather than S. If $B = 1$, then $C = 1$ if either $A = 1$ or $S = 1$; if $B = 0$, then $C = 1$ only if $A = 1$ and $S = 0$."*

Ambiguities (Example 3, Scenario 2)

Halpern (2016b)

> *"But now consider a slightly different story. This time, we view B as the switch, rather than S. If B = 1, then C = 1 if either A = 1 or S = 1; if B = 0, then C = 1 only if A = 1 and S = 0."*

$$C := (B \land (A \lor S)) \lor (\neg B \land (A \land \neg S))$$

Ambiguities (Example 3, Scenario 2)

Halpern (2016b)

> *"But now consider a slightly different story. This time, we view B as the switch, rather than S. If $B = 1$, then $C = 1$ if either $A = 1$ or $S = 1$; if $B = 0$, then $C = 1$ only if $A = 1$ and $S = 0$."*

$$C := (B \wedge (A \vee S)) \vee (\neg B \wedge (A \wedge \neg S))$$

What is the actual cause of C, intuitively?

- when B, then A or S or both
- when $\neg B$, then A and $\neg S$

Structural equation models are ambiguous

- It must be the case that some information of these informal scenarios is not expressed by the structural equation model.
- This information does not affect the possible causal worlds.
- This information affects the answer to actual causation problems!

Structural equation models are ambiguous

- It must be the case that some information of these informal scenarios is not expressed by the structural equation model.
- This information does not affect the possible causal worlds.
- This information affects the answer to actual causation problems!

What kind of information is that? Let's go back to the example.

The extra information

> *"There are four endogenous binary variables, A, B, C, and S, taking values 1 (on) and 0 (off). Intuitively, A and B are supposed to be alternative causes of C, and S acts as a switch. If S = 0, the causal route from A to C is active and that from B to C is dead; . . . "*

The extra information:

- separate causal mechanisms
- causes versus switches for causal mechanisms
- causal processes
- causal mechanisms can be alive or dead

dead $\sim$preempted

Solutions for the ambiguity

- Halpern's solution is a KR methodology :

  > *"what turns out to arguable be the best way to do the disambiguation is to add [. . .] extra variables, which [. . .] capture the* **mechanism of causality***".*
  >
  > *"But all this talk of mechanisms [. . .] suggests that the mechanism should be part of the model".*

- The approach of our paper:
  - Develop a formal language in which the missing information can be expressed.
  - Definitions of actual causation that exploit the extra information.
  - A formalization of the causal *process*
  - No counterfactual definitions (white-box system!)
  - Main goal: a framework to study various definitions of AC in

The idea

- We see separate causal mechanisms
- Some sets of conditions trigger the causal mechanism
- Other conditions could preempt the causal mechanism if not true; they enable/disable the mechanism.

Information about this strongly influences our idea of actual causation.

Syntax: Causal theories

Definition

A causal theory is a a set of causal mechanisms.

Definition

A causal mechanism, or causal law, is an expression of the form

$$A \leftarrow T \,||\, P$$

where A is a literal, T and S sequences of literals

- A literal of T is called a triggering condition of the causal mechanism.
- A literal of P is called an enabling condition of the causal mechanism.

Example 1

- Scenario 1:

$$\begin{aligned} Dead &\leftarrow Arsenic \\ Dead &\leftarrow Strychnine. \end{aligned}$$

- Scenario 2:

$$\begin{aligned} Dead &\leftarrow Arsenic \\ Dead &\leftarrow Strychnine || \neg Arsenic. \end{aligned}$$

Example 3

- Scenario 1:

$$\left\{\begin{array}{l} C \leftarrow A \,||\, \neg S \\ C \leftarrow B \,||\, S \end{array}\right\}$$

- Scenario 2:

$$\left\{\begin{array}{l} C \leftarrow A \,||\, B \\ C \leftarrow S \,||\, B \\ C \leftarrow A, S \,||\, \neg B \end{array}\right\}$$

- We made the information explicit that was available in Halperns informal domain description.

Semantics: possible causal processes

- A possible causal world semantics is not refined enough.
- The formal semantics specifies, for a causal theory Δ:
    - the possible causal processes of Δ
    - the possible causal world that each process leads to.
- How to formalize the causal process?
    - a causal process $\sim$ a dependency graph of the true literals, where edges labeled with:
        1. mechanisms that fire,
        2. role of the literal in the mechanism

Another example: double preemption

Hall (2004)

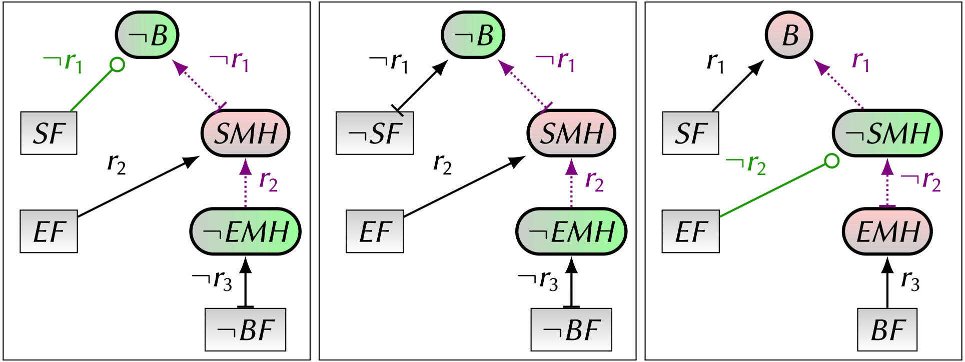
> *Suzy fires a missile (SuzyF) to bomb a target (B); Enemy fires a missile (EnemyF) to hit Suzy's missile (SuzyMH) and Billy fires a missile (BillyF) to hit Enemy's missile (EnemyMH).*

Theory:

$$\left\{ \begin{array}{l} B \leftarrow SuzyF \,||\, \neg SuzyMH \\ SuzyMH \leftarrow EnemyF \,||\, \neg EnemyMH \\ EnemyMH \leftarrow BillyF \,|| \end{array} \right\}$$

Another example: double preemption

$$\left\{ \begin{array}{l} B \leftarrow SuzyF \,||\, \neg SuzyMH \\ SuzyMH \leftarrow EnemyF \,||\, \neg EnemyMH \\ EnemyMH \leftarrow BillyF \,||\, \end{array} \right\}$$



Derived concepts and properties

Derived concepts:

- An actual possible causal process induces a unique possible causal world
  - The possible causal process semantics is more refined than the possible world semantics.
- In a possible world, a causal mechanism can be:
  - firing
  - triggered but preempted
  - non-triggered

Derived concepts and properties

Derived concepts:

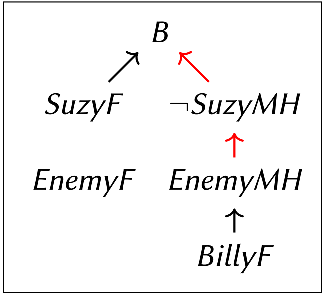
- An actual possible causal process induces a unique possible causal world
  - The possible causal process semantics is more refined than the possible world semantics.
- In a possible world, a causal mechanism can be:
  - firing
  - triggered but preempted
  - non-triggered

Some derived properties:

- All processes in the same exogeneous state cause the same possible world (confluence property)
- The possible causal worlds of Δ are the causal worlds of the structural equation model $\mathit{Completion}(\Delta)$

Definitions of actual causation

The possible causal process is a detailed explanation of the world.



B

$SuzyF$ $\neg SuzyMH$

$EnemyF$ $EnemyMH$

$BillyF$

- x is an influence of y in possible causal process $\mathcal{P}$
- x is an actual P-cause of y
- x is an actual DP-cause of y
- …

Example

Suzy and Billy both pick up rocks and throw them at a bottle....

- In this domain: many variations
- Common causal information:

$$\begin{aligned} \mathit{Shatter} &\leftarrow \mathit{SuzyThrows} \,|| \\ \mathit{Shatter} &\leftarrow \mathit{BillyThrows} \,|| \end{aligned}$$

- other information (who throws, who hits first, ...) is information about the actual causal process (not about the causal domain)

$$\begin{aligned} \mathit{Shatter} &:= \mathit{SuzyHits} \vee \mathit{BillyHits} \\ \mathit{SuzyHits} &:= \mathit{SuzyThrows} \\ \mathit{BillyHits} &:= \mathit{BillyThrows} \wedge \neg \mathit{SuzyThrows} \end{aligned}$$

Conclusions

- A study of several sorts of knowledge that are important for actual causation but are not or not well expressed in many causal languages.
- Logic equipped with a possible causal process semantics.
- Some fundamental aspects of causation: the confluence of causal processes and, paradoxically, a theorem explaining why many useful causation problems can be solved without modelling mechanisms and processes.
- A rich and flexible framework for defining several notions of actual causation.

Future work

- Relation with counterfactual definitions?
- Extending the logic: predicate logic, cyclic causal theories, . . .

Implementation on-line

`http://adams.cs.kuleuven.be/idp/server.html?chapter=intro/11-AC`

- An on-line implementation of many of the examples in the paper
- Using the knowledge base system IDP