Principles of
Knowledge Representation and Reasoning

Reasoning about Actual Causality

Bernhard Nebel, Felix Lindner, and Thorsten Engesser

April 17, 2018

UNI
FREIBURG

1 HP Definitions of Actual Causality, and Normality

Rock Example (Intro)

Example (Throwing Rock at Bottle)

Suzy and Billy both throw rocks at a bottle, but Suzy's hits the
bottle, and Billy's doesn't (although it would have hit had Suzy's
not hit first). The bottle shatters. Who caused the bottle to
shatter?

UNI
FREIBURG

Rock Example (Model)

- Model M involves five (boolean) endogenous variables ST (Suzy throws), BT (Billy throws), SH (Suzy's rock hits the bottle), BH (Billy's rock hits the bottle), BS (bottle shatters).

- The exogenous variable U ranges over pairs of boolean values determining who throws and who does not.

- Structural equations:
    - $ST := U = (1,0) \vee U = (1,1)$
    - $BT := U = (0,1) \vee U = (1,1)$
    - $SH := ST = 1$
    - $BH := BT = 1 \wedge SH = 0$
    - $BS := SH = 1 \vee BH = 1$

- In $(M, (1,1))$, neither ST nor BT are but-for causes of BS. But intuitively, we want ST be the cause of BS but not BT.

UNI FREIBURG

The Template of HP-Definitions

Definition (Actual Cause)

$\vec{X} = \vec{x}$ is an actual cause of φ in the causal setting $(M, \vec{u})$ iff

- **AC1**: $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \varphi$
- **AC2**: see next slides
- **AC3**: $\vec{X}$ is minimal, i.e., there is no strict subset $\vec{X}'$ of $\vec{X}$, s.th. $\vec{X}' = \vec{x}'$ satisfies conditions AC1 and AC2, where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}'$.

Original HP Definition

Definition (Original HP)

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

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

- **AC2(b^o)**: If $\vec{z}^*$ is such that $(M, \vec{u}) \models \vec{Z} = \vec{z}^*$, then for all subsets $\vec{Z}'$ of $\vec{Z} - \vec{X}$, we have

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

UNI
FREIBURG

Rock Example: Suzy is a Cause

- Is ST a cause of BS in situation $(M, (1, 1))$? Yes.
    - **AC1**:
        - $(M, (1, 1)) \models ST = 1$ and $(M, (1, 1)) \models BS = 1$ ☺
    - **AC2**:
        - Guess $\vec{Z} = \{ST, SH, BH, BS\}, \vec{W} = \{BT\}, w = 0$
        - **(a)**: $(M, (1, 1)) \models [ST \leftarrow 0, BT \leftarrow 0]\neg BS$ ☺
        - **(b^o)**: $(M, (1, 1)) \models [ST \leftarrow 1, BT \leftarrow 0]BS$,
          $(M, (1, 1)) \models [ST \leftarrow 1, BT \leftarrow 0, SH \leftarrow 1]BS$,
          $(M, (1, 1)) \models [ST \leftarrow 1, BT \leftarrow 0, BH \leftarrow 0]BS$,
          $(M, (1, 1)) \models [ST \leftarrow 1, BT \leftarrow 0, BS \leftarrow 1]BS$,
          $(M, (1, 1)) \models [ST \leftarrow 1, BT \leftarrow 0, SH \leftarrow 1, BH \leftarrow 0]BS$,
          $(M, (1, 1)) \models [ST \leftarrow 1, BT \leftarrow 0, SH \leftarrow 1, BS \leftarrow 1]BS$,
          $(M, (1, 1)) \models [ST \leftarrow 1, BT \leftarrow 0, BH \leftarrow 0, BS \leftarrow 1]BS$,
          $(M, (1, 1)) \models [ST \leftarrow 1, BT \leftarrow 0, SH \leftarrow 1, BH \leftarrow 0, BS \leftarrow 1]BS$
          ☺
    - **AC3**: ST is a singleton ☺

Nebel, Lindner, Engesser – KR&R

UNI FREIBURG

Rock Example: Billy is no Cause

- Is BT a cause of BS in situation $(M, (1, 1))$? No.
  - **AC1**:
    - $(M, (1, 1)) \models BT = 1$ and $(M, (1, 1)) \models BS = 1$ ☺
  - **AC2**:
    - Now we have to show that there is no partition by exhaustingly searching for it and finally failing. For example, consider $\vec{Z} = \{BT, SH, BH, BS\}$, $\vec{W} = \{ST\}$, $w = 0$
    - **(a)**: $(M, (1, 1)) \models [BT \leftarrow 0, ST \leftarrow 0]\neg BS$ ☺
    - **(b^o)**: $(M, (1, 1)) \models [BT \leftarrow 1, ST \leftarrow 0, BH \leftarrow 0]\neg BS$ ☹
    - Next try: $\vec{Z} = \{BT, SH, BS\}$, $\vec{W} = \{ST, BH\}$, $w = (0, 1)$
    - **(a)**: $(M, (1, 1)) \models [BT \leftarrow 0, ST \leftarrow 0, BH \leftarrow 0]\neg BS$ ☺, but then same problem as before for **(b^0)**. Otherwise $(M, (1, 1)) \models [BT \leftarrow 0, ST \leftarrow 0, BH \leftarrow 1]BS$ ☹.
  - **AC3**: BT is a singleton ☺

UNI FREIBURG

Witness

Definition (Witness)

The tuple $(\vec{W}, \vec{w}, \vec{x}')$ in condition AC2 of the HP definitions of causality are said to be a witness to the fact that $\vec{X} = \vec{x}$ is a cause of φ. The witness $(\emptyset, \emptyset, \vec{x}')$ denotes the special case that $\vec{W} = \emptyset$.

Example (Witness of Suzy causing the Bottle's Shattering)

$(\{BT\}, 0, 0)$

Shooting Example (Model)

Example (Shooting)

A prisoner dies either if A loads B's gun and B shoots, or if C loads and shoots his gun.

- Endogeneous variables D (prisoner's death), A (A loads B's gun), B (B shoots), C (C loads and shoots).
- $D := (A \wedge B) \vee C$, values of A, B, C are determined by one exogeneous variable U in the obvious way.
- In situation $(M, (1, 0, 1))$, A loads B's gun, B does not shoot, but C shoots (consequently, the prisoner dies).
- Is A is a cause for D in $(M, (1, 0, 1))$?

Shooting Example: A is a Cause of D

- Is A is a cause for D in $(M, (1, 0, 1))$? Yes.
  - Consider witness $(\{B, C\}, (1, 0), 0)$, i.e., set $\vec{Z} = \{A, D\}$, $\vec{W} = \{B, C\}$, and $\vec{w} = (1, 0)$
  - **AC2(a)** $(M, (1, 0, 1)) \models [A \leftarrow 0, B \leftarrow 1, C \leftarrow 0]D = 0$ ☺
  - **AC2(b^o)**: $(M, (1, 0, 1)) \models [A \leftarrow 1, B \leftarrow 1, C \leftarrow 0]D = 1$, $(M, (1, 0, 1)) \models [A \leftarrow 1, B \leftarrow 1, C \leftarrow 0, D \leftarrow 1]D = 1$ ☺
- The sufficiency conditions seems to be too weak.

Updated HP-Definition

Definition (Updated HP)

- **AC2(a)** same as original HP definition
- **AC2(b^u)** If $\vec{z}^*$ is such that $(M, \vec{u}) \models \vec{Z} = \vec{z}^*$, then for all subsets $\vec{W}'$ of $\vec{W}$ and subsets $\vec{Z}'$ of $\vec{Z} - \vec{X}$, we have

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

- According to updated HP definition, φ must hold even if only some of the values in $\vec{W}$ are set to w.
- In the shooting example and under the chosen $\vec{Z}$, $\vec{W}$, w, we get $(M, \vec{u}) \not\models [A \leftarrow 1, C \leftarrow 0]\neg(D = 1)$, so A's loading the gun was not sufficient for D's death, and hence, A did not cause D according to the updated HP definition.

UNI FREIBURG

Modified HP Definition

Definition (Modified HP)

- **AC2(a^m)** 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

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

- Here, the idea is that all that counts are the values the variables had in the situation to be analysed. So, this definition just asks if $\vec{X}$ is a but-for cause given we fix some of the variables to their actual values.

- No need for an extra sufficiency condition: We already know that φ holds when the variables were not changed.

Modified HP Definition: Some Notes

- Computationally simpler than the original and updated definitions.
- Solves the problems both in the Rock example, witness $(\{BH\}, 0, 0)$, and in the Shooting example (no witness for A).
- Suffers from similar problems as but-for causality in disjunctive forest fire. But: Considering Disjunctive Causes is an option! $L = 1 \vee MD = 1$ being a cause of FF just means that the fact that at least one of $L = 1$, $MD = 1$ holds is the cause of FF.

UNI
FREIBURG

Relationships (without proofs)

Theorem (see Halpern, Proposition 2.2.2)

If $X = x$ is a but-for cause of $Y = y$ in $(M, \vec{u})$, then $X = x$ is a cause of $Y = y$ according to all three variants of the HP definition.

Theorem (see Halpern, Proposition 2.2.3)

- *If $X = x$ is part of a cause of φ in $(M, \vec{u})$ according to the modified HP definition, then $X = x$ is a cause of φ in $(M, \vec{u})$ according to the original and the updated HP definition.*

- *If $X = x$ is part of a cause of φ in $(M, \vec{u})$ according to the updated HP definition, then $X = x$ is a cause of φ in $(M, \vec{u})$ according to the original HP definition.*

UNI
FREIBURG

Normality

Example (Normality, Knobe & 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. On Monday morning, one of the administrative assistants
encounters Professor Smith walking past the receptionist's desk.
Both take pens. Later that day, the receptionist needs to take an
important message, but she has a problem: There are no pens
left on her desk.

- Who is the cause of there not being pens?
- Kahnemann and Miller:"an event is more likely to be
  undone by altering exceptional than route aspects of the
  causal chain that led to it".

Nebel, Lindner, Engesser – KR&R

Extended Causal Model

Definition (Extended Causal Model)

An extended causal model is a tuple $M = (\mathcal{S}, \mathcal{F}, \succeq)$, where $(\mathcal{S}, \mathcal{F})$ is a causal model, and $\succeq$ is a partial preorder (reflexive, transitive) on worlds.

Definition (World)

In a recursive extended causal model M, a context $\vec{u}$ and interventions $\vec{X} = \vec{x}$ together determine a world $s_{\vec{X}=\vec{x},\vec{u}}$, viz., a complete assignment of values to all variables in M.

UNI FREIBURG

Normality Example: Extended Model

- Exogeneous variable U determines the truth of PS (Prof. Smith takes a pen) and PA (administrative assistant takes a pen).
  - $PS := U = (1,0) \lor U = (1,1)$, $AP := U = (0,1) \lor U = (1,1)$
- Variable NP is true in case both PS and PA are true.
  - $NP := PS \land PA$
- Relevant part of $\succeq$ for context $\vec{u} = (1,1)$:
  - $s_{PS=0,\vec{u}} \succ s_{\vec{u}}$: The world in which Smith takes no pen and the assistant does, is more normal than the world in which both take a pen.
  - $s_{\vec{u}} \succ s_{PA=0,\vec{u}}$: The world in which both take a pen, is more normal than the world in which Smith takes a pen and the assitant does not.

UNI FREIBURG

Extended Modified HP Definition

Definition (Extended Modified HP Definition)

- **AC2⁺(aᵐ)** 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

$$s_{\vec{X} = \vec{x}', \vec{W} = \vec{w}^*, \vec{u}} \succeq s_{\vec{u}}$$

and

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

- So, if we have to make a situation more untypical in order to prove some $\vec{X} = \vec{x}$ a cause, then it is not a cause.
- The original and updated HP definitions can be extended in a similar way.

UNI
FREIBURG

Normality Example: It was Prof. Smith!

- In $(M, (1,1))$, $PS = 1$ is a cause of $NP = 1$ according to the extended modified HP definition:
    - **AC1**: $(M, (1,1)) \models PS = 1 \land NP = 1$ 😊
    - **AC2$^+$(a^m)**: Consider witness $(\emptyset, \emptyset, 0)$:
        - $s_{PS=0,\vec{u}} \succeq s_{\vec{u}}$ ☺
        - $(M, \vec{u}) \models [PS \leftarrow 0] \neg (NP = 1)$ 😊
    - **AC3**: $PS = 1$ is a singleton 😊
- But $PA = 1$ is not a cause of $NP = 1$:
    - **AC1**: $(M, (1,1)) \models PA = 1 \land NP = 1$ 😊
    - **AC2$^+$(a^m)**: $s_{PA=0,\vec{u}} \not\succeq s_{\vec{u}}$ ☹

UNI FREIBURG

Outlook

- Responsibility & Blame
- Explanation

Nebel, Lindner, Engesser – KR&R

2 Literature

Literature I

Pearl, J., Mackenzie, D.
The Book of WHY – The New Science of Cause and Effect,
Basic Books, 2018.

Halpern, J. Y.
Actual Causality,
MIT Press, 2016.