Multi-Agent Systems

Reasoning about Actual Causality

Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Bernhard Nebel, Felix Lindner, and Thorsten Engesser

April 17, 2018

Motivation

- AI systems are used or are about to be used in domains that potentially affect people's life significantly: Finance, Law, Health etc.
- According to The European Union General Data Protection Regulation, everyone has the right to obtain an explanation of the decision reached [...] and to challenge the decision.
- In AI, there is currently a huge interest in so-called Explainable AI (XAI), i.e., the design and analysis of systems that are able to explain their decisions to humans.
- In an Multi-Agent System, deriving causal explanations can cleafify which agents' actions resulted in some proposition to become true, i.e., assigning responsibility and, maybe, blame.

Pearl’s Ladder of Causation

Causation

- Everyone who has ever taken a statistics class has probably been told that correlation is not causation. But what is causation then?
- We will first learn about Judea Pearl's Ladder of Causation distinguishing three reasoning modes: Association (Seeing), Intervention (Doing), and Introspection (Imagining).
- We will then study Judea Pearl's and Joseph Halpern's attempts to define causality and related concepts based on causal models [1, 2].

Association: Seeing

- Answers questions like “What if I see ...”?, “How would seeing X change my belief in Y?”
- E.g.: Seeing a high number on the thermometer makes me believe it is sunny outside. Seeing features X, Y, Z in an image makes the AI believe that there is a cat on the picture.
- Correlation between variables.

Intervention: Doing

- Answers questions like "What if I do …", "What would Y be if I do X?", "How can I make Y happen?"
- E.g.: Taking an aspirin will cure my headache. But, heating the thermometer will not make the sun shine.
- This type of reasoning requires to disentangle otherwise correlated variables.

Introspection: Imagining

- Answers questions like “What if I had (not) done ...?”, “Was it X that caused Y?”, “What if X had not occured?”
- Being able to answer such question is a prerequisite for AI systems to reason about:
  - Regret: Would things have turned out better if I had acted otherwise?
  - Responsibility: To what extend was it my action that caused X?
  - Blame: Could/Should I have known that my action will cause X?
- This type of reasoning requires to fix some variables to the value they had in a particular situation while changing the values of other variables, i.e., considering counterfactual worlds.

Causal Models

Definition (Causal Model)

A causal model M is a pair $(\mathcal{S},\mathcal{F})$, where

- $\mathcal{S} = (\mathcal{U},\mathcal{V},\mathcal{R})$ is a signature, which explicitly lists the exogeneous variables $\mathcal{U}$, the endogeneous variables $\mathcal{V}$, and associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a non-empty set $\mathcal{R}(Y)$ of possible values for Y,
- $\mathcal{F}$ associates one structural equation F_X to each endogeneous variable $X \in \mathcal{V}$: $F_X : \mathcal{R}(Z_1) \times \ldots \times \mathcal{R}(Z_{|\mathcal{U}\cup\mathcal{V}|-1}) \to \mathcal{R}(X)$ for all $Z_i \in \mathcal{U} \cup \mathcal{V} - \{X\}$

Terminology

- Model M: Specification of the available variables (exogeneous and endogeneous) and their structural relationships (via structural equations).
- Context $\vec{u}$: An assignment of values to the exogeneous variables. (From this assignment, the values of the endogeneous variables can be deterministically determined).
- Setting $(M, \vec{u})$: A pair of a model and a context determines a setting. In a setting, every variable in the model has got a value.

Intervention

Definition (Intervention)

An intervention sets the value of some endogeneous variable X to a value x in a causal model $M = (\mathcal{S}, \mathcal{F})$ resulting in a new causal model $M_{X \leftarrow x} = (\mathcal{S}, \mathcal{F}_{X \leftarrow x})$, where $\mathcal{F}_{X \leftarrow x}$ results from replacing the structural equation for X in $\mathcal{F}$ by $X = x$ and leaving the remaining equations untouched.

- Interventions enable counterfactual reasoning by setting values different from actual values thereby overriding structural equations.

Independence and Recursiveness I

Definition (Independence)

Endogeneous variable Y is independent of endogeneous variable X in a setting $(M,\vec{u})$ iff for all settings $\vec{z}$ of the endogeneous variables other than X and Y, and all values x,x' of X, $F_Y(x,\vec{z},\vec{u}) = F_Y(x',\vec{z},\vec{u})$ holds.

Definition (Recursive Model)

A model M is recursive iff for each context $\vec{u}$, there is a partial order $\preceq_{\vec{u}}$ (reflexive, anti-symmetric, transitive) of the endogeneous variables, such that unless $X \preceq_{\vec{u}} Y$, Y is independent of X in $(M,\vec{u})$.

Independence and Recursiveness II

- Independence may vary depending on context $\vec{u}$. Consider $M = (\mathcal{S}, \mathcal{F})$:
  - $\mathcal{S} = (\{C\}, \{X, Y\}, \{C \mapsto \{0,1\}, X \mapsto \{0,1\}, Y \mapsto \{0,1\}\})$
  - $\mathcal{F} = \{X := (C = 1) \wedge (Y = 1), Y := (C = 1) \vee (X = 1)\}$[1]
- Case $\vec{u} = (0)$: X is independent of Y, Y depends on X.
- Case $\vec{u} = (1)$: X depends on Y, Y is independent of X.

[1] We here abuse notation a bit.

Independence and Recursiveness III

- For a recursive model M and context $\vec{u}$, the value of all endogeneous variables can be determined deterministically:
  - First, determine values of variables that depend only on $\vec{u}$ (first level).
  - Second, determine values of variables that depend only on $\vec{u}$ and first-level variables (second level).
  - ...
- In everything that follows, "causal model" will always mean "recursive causal model".

Language of Causality: Syntax

- Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$. A causal formula over $\mathcal{S}$ is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where
  - φ is a boolean combination (using $\wedge, \vee, \neg, \rightarrow$) of primitive events (of the form $X = x$), and
  - $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and
  - $y_i \in \mathcal{R}(Y_i)$.
- Common abbrevation: $[\vec{Y} \leftarrow \vec{y}]\varphi$
- Case $k = 0$: $[]\varphi$ is also just written as φ

Language of Causality: Semantics

- Truth of a causal formula is validated relative to a causal model M and a context $\vec{u}$.
- $(M,\vec{u}) \models X = x$ iff the value of X is x once the exogeneous variables are set to $\vec{u}$.
- $(M,\vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ iff $(M_{\vec{Y} \leftarrow \vec{y}},\vec{u}) \models \varphi$
- Boolean combinations validated as usual: $(M,\vec{u}) \models \varphi \wedge \psi$ iff $(M,\vec{u}) \models \varphi$ and $(M,\vec{u}) \models \psi$ etc.

But-For Cause

Definition (Cause according to Hume)

"We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed."

Definition (But-For Cause)

$X = x$ is a but-for cause of φ in $(M,\vec{u})$ iff

- $(M,\vec{u}) \models (X = x) \land \varphi$, and
- there exists some x', s.th. $(M,\vec{u}) \models [X \leftarrow x']\neg\varphi$

Forest Fire: Conjunctive

Example (Conjunctive Forest Fire)

- Consider M^c with exogeneous variable U, and endogeneous variables L (lightning), MD (dropped match), FF (forest fire), s.th. $\mathcal{R}(U) = \{(0,0),(0,1),(1,0),(1,1)\}$, $\mathcal{R}(L) = \mathcal{R}(MD) = \mathcal{R}(FF) = \{0,1\}$, and $L := U = (1,0) \vee U = (1,1)$, $MD := U = (0,1) \vee U = (1,1)$, $FF := L = 1 \wedge MD = 1$.
- Did the lightning (L) cause the forest fire (FF) in setting $M,(1,1)$? Check for but-for cause:
  - $(M,(1,1)) \models L = 1 \wedge FF = 1$
  - $(M,(1,1)) \models [L \leftarrow 0]\neg FF$
- Answer: Yes.

Forest Fire: Disjunctive

Example (Disjunctive Forest Fire)

- Consider M^d, which differs from M^c only in the structural equation for FF, viz., $FF := L = 1 \vee MD = 1$.
- Again: Did the lightning (L) cause the forest fire (FF) in setting $M, (1,1)$? Check for but-for cause:
  - $(M, (1,1)) \models L = 1 \wedge FF = 1$
  - $(M, (1,1)) \not\models [L \leftarrow 0]\neg FF$
- Answer: No.
- Using the same reasoning, MD also is not a cause according to the but-for definition of causality.
- (But $L \vee MD$ is.)

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?

Rock Example (Model)

- Model M involves five (boolean) endogeneous 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 exogeneous 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.

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($\mathbf{b}^o$)**: 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

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

- AC2(a): X is necessary to make φ happen (in some related world).
- AC2(b): X is sufficient to trigger actual causal path Z to φ.

Rock Example: Suzy is a Cause

- Is ST a cause of BS in setting $(M,(1,1))$? 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 ☺

Rock Example: Billy is no Cause

- Is BT a cause of BS in setting $(M,(1,1))$? 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)**: $(M,(1,1)) \models [BT \leftarrow 0, ST \leftarrow 0, BH \leftarrow 1]BS$ ☹.
  - **AC3**: BT is a singleton ☺

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).
- $D := (A \land B) \lor C$, 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($\mathbf{b}^u$)** 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.

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 setting 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.

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.*

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".

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.

Normality Example: Extended Model

- Exogeneous variable U determines the truth of PS (Prof. Smith takes a pen) and PA (administrative assistant takes a pen).
  - $PS := U = (1,0) \vee U = (1,1)$, $AP := U = (0,1) \vee U = (1,1)$
- Variable NP is true in case both PS and PA are true.
  - $NP := PS \wedge PA$
- 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.

Extended Modified HP Definition

Definition (Extended Modified HP Definition)

- **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

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

and

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

- So, if we have to make a setting 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.

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**: $(M,(1,1)) \models PS = 1 \wedge NP = 1$ ☺
  - **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**: $(M,(1,1)) \models PA = 1 \wedge NP = 1$ ☺
  - **AC2⁺(a^m)**: $s_{PA=0,\vec{u}} \not\succeq s_{\vec{u}}$ ☹

Outlook

- Responsibility & Blame

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.