Principles of Knowledge Representation and Reasoning

Reasoning about Actual Causality

Bernhard Nebel, Felix Lindner, and Thorsten Engesser

April 17, 2018

Motivation

- AI systems are used or are about to be used in many 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.
- That's a perfect reason (among others) to study causal reasoning as a means to come up with answers to Why-questions, i.e., explanations.

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

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 $X \in \mathcal{V}$: $F_X : \mathcal{R}(Z_1) \times \ldots \times \mathcal{R}(Z_{|\mathcal{U} \cup \mathcal{V}|-1}) \rightarrow \mathcal{R}(X)$ for all $Z_i \in \mathcal{U} \cup \mathcal{V} - \{X\}$

UNI FREIBURG

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).
- Situation $(M, \vec{u})$: A pair of a model and a context determines a situation. In a situation, 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 $M = (\mathcal{S}, \mathcal{F})$:
  - $\mathcal{S} = (\{C\}, \{X, Y\}, \{C \mapsto \{0,1\}, X \mapsto \{0,1\}, Y \mapsto \{0,1\}\})$
  - $\mathcal{F} = \{X := (C = 1) \land (Y = 1), Y := (C = 1) \lor (X = 1)\}$[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

- $(M, \vec{u}) \models (X = x) \wedge \varphi$, 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 situation $M,(1,1)$? 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.

UNI FREIBURG

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 situation $M,(1,1)$? 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.)

UNI FREIBURG

Outlook

- Halpern-Pearl-Definitions of Causality
- Normality, Responsibility, and Blame
- Explanation

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.