Logical Foundations for Actual Causality

Shonan Seminar No.139

Vitaliy Batusov[1] Mikhail Soutchanski [2]

[1]York University, Toronto

[2]Ryerson University, Toronto

June 25, 2019

Overview

1. Motivation
   - Problems with HP
2. Proposal
   - Situation Calculus
   - Achievement Causes
   - Achievement Causal Chain
   - Maintenance Causes
   - Actual Cause
3. Relationship with Halpern-Pearl
4. Conclusions

We envision that the following research questions will be addressed in the course of the seminar:

- What are current research activities in the area of causal reasoning, and what application scenarios are considered? What new and promising application areas of causal reasoning can be identified? In the light of changing paradigms of computing, how will causal reasoning have to change? For instance, what is the impact of the autonomy of cyber-physical systems on the notion of causality? What impact does emergent behavior of large collections of computing devices have on causality? Can causality analysis help in explaining the result of a program, for instance, the decisions of deep neural networks? How to generate useful explanations?
- **How to characterize causality? Is there a better way to design “good” definitions of causality than relying on the trial-and-error scheme assessing candidate definitions on a host of textbook examples?**
- What can causal reasoning on computing systems, and in social context — for instance in litigation, tort law, or economy — learn from each other?
- How can causal reasoning be applied to security and privacy properties, e.g., to determine the actors responsible for information leakage?
- What calculi and tools are available to support causal reasoning? For which type of tools is there a demand, and what are the desiderata for such tools?
- How to scale causal analysis, other than statistical approaches, to real-world applications? How do causal analysis and abstraction compose? How to design systems for accountability, in the sense that in the case of a system failure, the causes can be determined automatically?
- Is there a compendium of open or unsatisfactorily solved problems?

Actual Cause

- Long-standing philosophical problem
- **Token-level** (actual) vs type-level (general) causality
- Actual cause: *A thing which* **has caused** *another thing*.
- Given a narrative and a statement ϕ that holds true at the end, how do we separate actual causes of ϕ from irrelevant events?
- Converting intuition into a formal definiton is not trivial
- Dominant approach in AI: *counterfactul analysis* in *systems of structural equations* (Pearl, Halpern, others)

Halpern and Pearl

- HP seems to work well **within its ontological limits**
- Or maybe not. See: "A quest for formal tools for reasoning about counterfactual causation" by Gössler, Stefani, Sokolsky

Problems of HP approach that we address

- Enduring conditions ("man is dead") vs. transitions ("man dies")
- Absence of event $\equiv$ presence of it opposite
- No objects, relationships, time, quantifiers in queries
- Distinct domains appear isomorphic; contradictory intuitions [Hopkins and Pearl, 2007, Glymour et al., 2010, Beckers and Vennekens, 2012]

- [Hopkins and Pearl, 2007]: Situation Calculus Causal Models
- Kept PWS and all issues related to counterfactuals
- Interventions realized via ignoring precondition axioms (!)
- Did not define actual cause

- Example-driven

Suzy and Billy

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

Starting From Scratch

- We want to search for causes in dynamical systems
- **Dynamical system**: a system that undergoes change of some kind
- Dynamic systems have states and transitions between them
- Examples: pendulum, animal populations, digital circuits
- Kinds of change: *discrete*, *continuous*, *hybrid*

Formalisms

- Need a suitable language to describe and analyse dynamical systems
- Various formalisms for different applications are available

- **Situation Calculus**: “system as a logical theory”

  Designed specifically to *formally* capture the phenomena of action, situation, and change

Regression: A Very Important Tool Indeed

- Regression: an extremely useful deductive tool
- Key idea: **reduce a query about some state to a logically equivalent query about a previous state**
- Proposition: *a formalism for causality must allow for regression*
- First proposed in 1960's as *retrosynthetic analysis* for organic chemistry by Elias James Corey (NP 1990 in chemistry)
- Introduced as *goal regression* in AI by Richard Waldinger (1975)
- Adapted into situation calculus by Raymond Reiter (1990's)

Situation Calculus

- System is described axiomatically in first-order logic
- Theory has three sorts: *action*, *situation*, *domain object*
- **actions**: symbols which trigger change
- **situations**: sequences of actions (world histories)
- **domain objects**: everything else (cats, cars, numbers, etc.)
- Predicates and functions describe properties of objects: $Cat(John)$
- Actions are used to construct situations: $feed(John)$ executed in the initial situation S_0 yields a new situation $do(feed(John), S_0)$
- Predicates/functions whose last argument is a situation are called *fluents*: $Happy(x, s)$ — "x is happy in situation s"
- Fluents are what changes from one situation to another. Thus, situations are a frame of reference, but fluents are the state.

Modelling Systems in Situation Calculus

- *Basic Action Theories* — Reiter (2001)
- To describe dynamics of fluent F, begin with *effect axioms*

$$\phi^{+}(\bar{x}, a, s) \rightarrow F(\bar{x}, do(a, s)) \qquad \text{(positive)}$$
$$\phi^{-}(\bar{x}, a, s) \rightarrow \neg F(\bar{x}, do(a, s)) \qquad \text{(negative)}$$

  Note: these are general causal rules (*type-level causality*)
- Example:

$$Cat(x) \wedge \neg Happy(x, \overbrace{s}^{\text{situation}}) \rightarrow Happy(x, \overbrace{do(feed(x), s)}^{\text{successor situation}})$$
$$Cat(x) \rightarrow \neg Happy(x, do(bathe(x), s))$$

Frame problem

- Want the theory to unambiguously describe what happens when an action is executed
- Things that change — effect axioms
- Things that simply carry over — ? (too many to list)

Reiter's solution

- Causal completeness assumption: *there is no other source of change to fluent F other than what is asserted in the effect axioms*
- Formally:

$$F(\bar{x}, s) \land \neg F(\bar{x}, do(a, s)) \rightarrow \phi^-(\bar{x}, a, s)$$
$$\neg F(\bar{x}, s) \land F(\bar{x}, do(a, s)) \rightarrow \phi^+(\bar{x}, a, s).$$

- Assuming that ϕ^+ and ϕ^- can never happen simultaneously, Reiter derives the *successor state axiom*

$$F(\bar{x}, do(a, s)) \leftrightarrow \phi^+(\bar{x}, a, s) \lor F(\bar{x}, s) \land \neg\phi^-(\bar{x}, a, s)$$

which is logically equivalent to the conjunction of the above axioms.

Successor State Axioms

- Example:

$$Happy(x, do(a, s)) \leftrightarrow a = feed(x) \wedge Cat(x) \wedge \neg Happy(x, s) \\ \vee Happy(x, s) \wedge \neg[a = bathe(x) \wedge Cat(x)]$$

- This concise form is the key feature of BATs
- Makes regression possible: given a query about a far-away situation, can transform it to equivalent query about S_0

Single-step Regression

- We use single-step regression operator ρ:
  $\rho[\varphi, \alpha]$ is obtained from φ by replacing each fluent atom by the RHS of its SSA while substituting α for action variable and simplifying, e.g.

$$High(x, do(a, s)) \leftrightarrow a = hi(x) \vee High(x, s) \wedge a \neq lo(x)$$
$$\rho[High(x, s), hi(c)] \text{ is } c \neq x \rightarrow High(x, s)$$

- $\mathcal{D} \models \forall s.\ \varphi(do(\alpha, s)) \leftrightarrow \rho[\varphi(s), \alpha]$

Causal Setting

Need a standard way of representing a causal scenario, e.g.

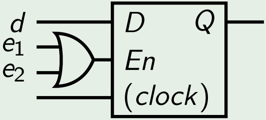
An arsonists drops a match in the forest and a lightning bolt strikes a tree. Either one of these events is sufficient to set the forest on fire. What caused the forest fire?

- Causes are actions, effects are FOL sentences (cf. HP)
- BAT captures dynamics, ground situation captures narrative

Definition

A (SC) *causal setting* is a triple $\langle \mathcal{D}, \sigma, \varphi(s) \rangle$ where $\mathcal{D}$ is a BAT, σ is a ground situation term such that $\mathcal{D} \models \mathit{executable}(\sigma)$, and $\varphi(s)$ is a SC formula uniform in s such that $\mathcal{D} \models \exists s(\mathit{executable}(s) \wedge \varphi(s))$.

Example (running)

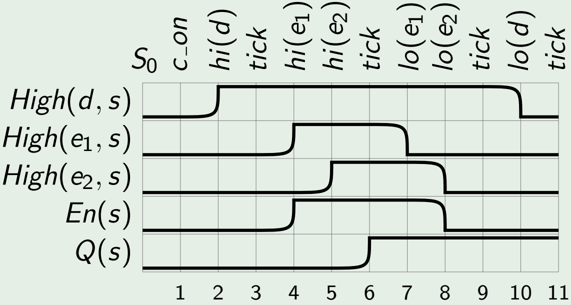


$$Poss(c\_on, s),$$
$$Poss(tick, s) \leftrightarrow ClockOn(s),$$
$$Poss(hi(x), s) \leftrightarrow \neg High(x, s),$$
$$Poss(lo(x), s) \leftrightarrow High(x, s),$$

$$\begin{aligned}
ClockOn(do(a, s)) &\leftrightarrow a = c_on \vee ClockOn(s),\\
High(x, do(a, s)) &\leftrightarrow a = hi(x) \vee High(x, s) \wedge a \neq lo(x),\\
En(do(a, s)) &\leftrightarrow a = hi(e_1) \vee a = hi(e_2) \vee\\
&\quad En(s) \wedge \neg[a = lo(e_1) \wedge \neg High(e_2, s)] \wedge\\
&\quad \neg[a = lo(e_2) \wedge \neg High(e_1, s)],\\
Q(do(a, s)) &\leftrightarrow [a = tick \wedge En(s) \wedge High(d, s)] \vee\\
&\quad Q(s) \wedge \neg[a = tick \wedge En(s) \wedge \neg High(d, s)].
\end{aligned}$$

Example

Causal setting: $\langle \mathcal{D}, \sigma, Q(s) \rangle$ where $\mathcal{D}$ is the BAT above and σ is $do([c\_on, hi(d), tick, hi(e_1), hi(e_2), tick, lo(e_1), lo(e_2), tick, lo(d), tick], S_0)$.

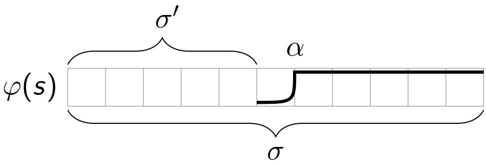


Causal roles include *achievement* and *maintenance*

Primary Achievement Cause

Definition

Let $\mathcal{C} = \langle \sigma, \phi \rangle$ be a setting. The action α executed in situation σ_α is an *achievement cause of* $\mathcal{C}$ iff $do(\alpha, \sigma_\alpha) \sqsubseteq \sigma$ and $\mathcal{D} \models \neg\phi(\sigma_\alpha) \land \forall s\,(do(\alpha, \sigma_\alpha) \sqsubseteq s \sqsubseteq \sigma \rightarrow \phi(s))$. We write $AC(\mathcal{C})$ to denote the situation term $do(\alpha, \sigma_\alpha)$ such that α executed in σ_α is an achievement cause of $\mathcal{C}$.



go back for an example

Achievement Causal Chain

- Captured "straw that broke camel's back", what about the rest?
- Recall: $\varphi(do(\alpha, \sigma')) \equiv \rho[\varphi(s), \alpha](\sigma')$ wrt $\mathcal{D}$
- Thus, if achievement condition is satisfied via $do(\alpha, \sigma')$, then $\rho[\varphi(s), \alpha]$ expresses a true, necessary, and sufficient condition in σ' for achieving $\varphi(s)$ via α
- Can apply last definition to $\langle \rho[\varphi(s), \alpha], \sigma' \rangle$ and repeat
- Must not overlook action preconditions!

Definition (Precursor)

Let $\mathcal{C} = \langle \sigma, \phi \rangle$ be a non-trivial setting and let $\sigma^\star = AC(\mathcal{C})$ such that $\sigma^\star = do(\alpha, \sigma_\alpha)$. The *achievement precursor of* $\sigma^\star$, denoted $Pre^A(\sigma^\star)$, is the setting $\langle \sigma_\alpha,\ \rho[\phi, \alpha] \wedge \Pi_\alpha \rangle$.

Causal Chain

Definition (Causal Chain)

Let $\mathcal{C} = \langle \sigma, \phi \rangle$ be a setting. An *achievement causal chain of* $\mathcal{C}$ is a sequence $\sigma_1, \sigma_2, \ldots$ such that $\sigma_1 = AC(\mathcal{C})$ and, for every σ_i with $i > 1$, $\sigma_i = AC(Pre^A(\sigma_{i-1}))$.

The ACC is a mathematical object which represents the achievement causes discovered starting from the given setting and going back through the precursors. Thus, causes appear in ACC in reverse chronological order.

(draw a picture)

$$\left\langle do([c\_on, hi(d), tick, hi(e_1), hi(e_2), \mathbf{tick}, lo(e_1), lo(e_2), tick, lo(d), tick], S_0), Q(s) \right\rangle$$

$$\Pi_{tick}(s) \qquad \rho[\cdot, tick]$$

$$\left\langle do([c\_on, hi(d), tick, \mathbf{hi(e_1)}, hi(e_2)], S_0), ClockOn(s) \wedge (\neg En(s) \vee High(d,s)) \wedge (En(s) \vee Q(s)) \right\rangle$$

$$\Pi_{hi}(e_1, s) \qquad \rho[\cdot, hi(e_1)]$$

$$\left\langle do([c\_on, \mathbf{hi(d)}, tick], S_0), \neg High(e_1, s) \wedge ClockOn(s) \wedge High(d,s) \right\rangle$$

$$\Pi_{hi}(d, s) \qquad \rho[\cdot, hi(d)]$$

$$\left\langle do([\mathbf{c\_on}], S_0), \neg High(d,s) \wedge \neg High(e_1, s) \wedge ClockOn(s) \right\rangle$$

$$\rho[\cdot, c\_on]$$

$$\left\langle S_0, \neg High(d,s) \wedge \neg High(e_1,s) \right\rangle$$

Maintenance Causes: General Considerations

- Achievement causal chain does not explain how an effect persists; searches backwards from primary AC
- An already-achieved effect cannot be achieved, but it can be destroyed
- Since it is not destroyed*, either it was not threatened, or the threats were neutralized by what we call *maintenance causes*
- Maintenance involves two actions: a *threat* and a *maintenance cause*
- Maintenance cause precedes the threat
- A false fact cannot be maintained or threatened
  (thus, the first action of a narrative is not a threat)
  (and neither is the primary achievement cause)
- A threat is capable of falsifying the effect (counterfactual test)

Maintenance Cause (cont.)

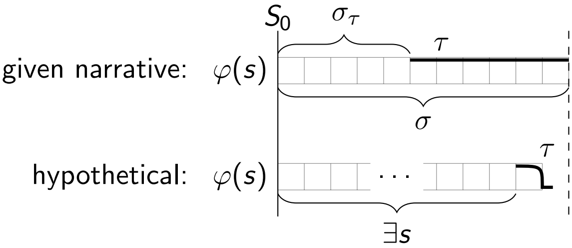
Definition (Threat)

Let $\mathcal{C} = \langle \sigma, \phi \rangle$. The action τ executed in σ_τ is a *threat to* $\mathcal{C}$ iff

1. $do(\tau, \sigma_\tau) \sqsubseteq \sigma$,
2. $\mathcal{D} \models \exists s\big(executable(do(\tau, s)) \land \phi(s) \land \neg\phi(do(\tau, s))\big)$, but
3. $\mathcal{D} \models \forall s(\sigma_\tau \sqsubseteq s \sqsubseteq \sigma \rightarrow \phi(s))$.

The *maintenance precursor* of $do(\tau, \sigma_\tau) \in Threats(\langle \sigma, \phi \rangle)$, denoted $Pre^M(do(\tau, \sigma_\tau))$, is the setting $\langle \sigma_\tau, \rho[\phi, \tau] \rangle$.

Maintenance Cause (cont.)



Definition

Let $\sigma^\star \in \mathit{Threats}(\mathcal{C})$. The action α executed in σ_α is a *maintenance cause of* $\mathcal{C}$ iff α executed in σ_α is the achievement cause of $\mathit{Pre}^M(\sigma^\star)$.

Example (altered scenario)

Consider the same BAT with an infinite set of signal constants $c_1, c_2, \ldots$

- Query $\varphi(s)$: $\exists x \exists y(x \neq y \wedge High(x, s) \wedge High(y, s))$
- Narrative σ: $do([hi(c_1), hi(c_2), hi(c_3), lo(c_1)], S_0)$.

$lo(c_1)$ is a threat. Yields a new causal setting $\langle do([hi(c_1), hi(c_2), hi(c_3)], S_0), \rho[\varphi(s), lo(c_1)]\rangle$, where $\rho[\varphi(s), lo(c_1)]$ is

$$\exists x \exists y(x \neq y \wedge High(x, s) \wedge High(y, s) \wedge x \neq c_1 \wedge y \neq c_1).$$

$hi(c_3)$ is an achievement cause here, so it is a maintenance cause in the original setting.

note the open domain and quantifiers

Actual Cause

- Preceding definitions do not capture interplay between achievement and maintenance
- Assumption: all causes of a descendant causal setting are equally relevant to ancestor setting

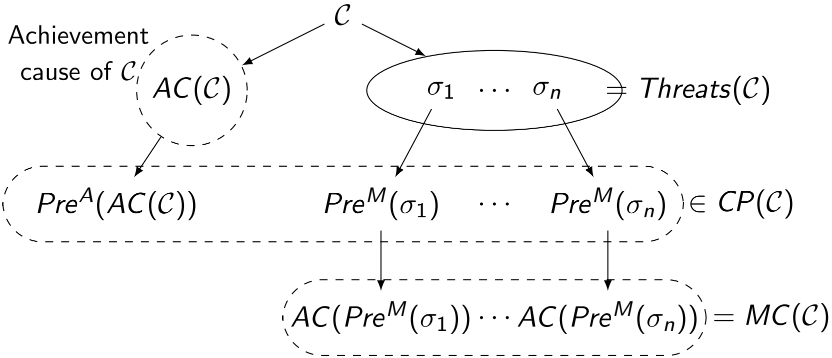
Definition (Actual Cause)

The set of *parental settings of* $\mathcal{C}$, denoted $PS(\mathcal{C})$, is the smallest set which contains $\mathcal{C}$ and, for each $\mathcal{C}' \in PS(\mathcal{C})$,

- $Pre^A(AC(\mathcal{C}')) \in PS(\mathcal{C})$;
- $Pre^M(\sigma^\star) \in PS(\mathcal{C})$ for each $\sigma^\star \in Threats(\mathcal{C}')$.

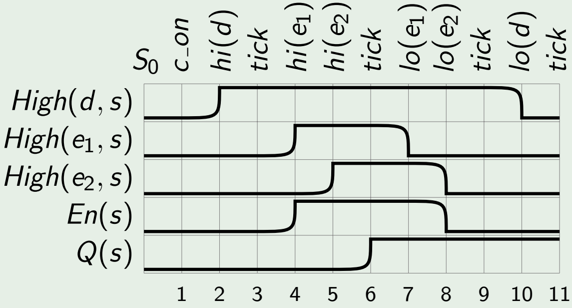
The action α executed in σ_α is an *actual cause of* $\mathcal{C}$ iff $do(\alpha, \sigma_\alpha) \in \{AC(\mathcal{C}') \mid \mathcal{C}' \in PS(\mathcal{C})\}$.

Inductive Tree of Actual Causality



Example (running example again)

The action $lo(e_2)$ is a non-trivial actual cause of $Q(s)$ discovered through a combination of two maintenance conditions (first threat *tick*, second threat $lo(d)$). It is not discoverable without the last definition.



Halpern-Pearl approach (very briefly)

- *Causal models* [Pearl, 1998]
- Multi-valued variables, e.g. the binary FF (forest is on fire, $true/false$)
- Structural equations, e.g. $FF := (MD = true) \wedge (L = true)$
- CM are acyclic (exists unique solution to equations)
- A language with semantics based on the unique solution
- *Interventions*: force values upon some of the variables, see what happens to "effect", e.g.

$$M, \bar{u} \models [MD \leftarrow false](FF = true)$$

- Latest version of defn. of AC (HP^m) is an incremental improvement [Halpern, 2015]

Formal relationship of HP and our approach

- Axiomatization schema turns an arbitrary HP to a SC causal setting
- Proved correctness of translation, cause correspondence

Theorem (Main result)

Let $(M, \bar{V}_U)$ be a HP causal setting and ϕ a HP query over M. Let $\mathcal{D}$ be a BAT obtained from $(M, \bar{V}_U)$. Let $X \in \mathcal{V}$ and $V_X \in \mathcal{R}(X)$.

1. *$(X = V_X)$ is a singleton cause of ϕ in $(M, \bar{V}_U)$ according to HP^m if and only if $get(X, V_X) \in \sigma$ appears in the achievement causal chain of $\langle \sigma, \hat{\phi}(s) \rangle$ for every ground situation term σ of $\mathcal{D}$ such that $\mathcal{D} \models interv_{\emptyset}(\sigma)$.*
2. *$(X = V_X)$ is a part of a cause of ϕ in $(M, \bar{V}_U)$ according to HP^m if and only if there exists a ground situation term σ of $\mathcal{D}$ such that $\mathcal{D} \models interv_{\emptyset}(\sigma)$ and $get(X, V_X) \in \sigma$ appears in the achievement causal chain of $\langle \sigma, \hat{\phi}(s) \rangle$.*

Illusory dillemmas

According to [Beckers and Vennekens, 2012], the following examples are isomorphic

Assassin poisons victim's coffee, victim drinks it and dies. If assassin had not poisoned the coffee, his backup would have, and victim would still have died. ([Hitchcock, 2007])

[Beckers and Vennekens, 2012]: intuitively, poisoning is a cause

An engineer is standing by a switch in the railroad track. A train approaches in the distance. She flips the switch, so that the train travels down the left-hand track instead of the right. Since the tracks re-converge up ahead, the train arrives at its destination all the same.

[Hall, 2000, Paul and Hall, 2013]

[Beckers and Vennekens, 2012]: intuitively, switching is not a cause

Conclusions

- Our proposal is based on a small set of plausible intuitions, yet is compatible with previous work built on completely different premises
- Rich ontology of SC takes "art" out of causal modelling [Halpern, 2016], e.g., transition $\neq$ condition
- Precondition axioms uncover a separate causal pathway ignored in previous work, allowing for better causal explanations
- Unrestricted FOL allows to analyze complex domains and "effects" with quantifiers

Current and Future Work

- Hybrid situation calculus — causes of continuous phenomena
- Absense of action as a cause
- Partially ordered or incomplete narratives
- Higher-level causes
- Attribution of causes to agents, responsibility and blame
- Causes across abstraction and refinement

Questions?

1. Motivation
   - Problems with HP
2. Proposal
   - Situation Calculus
   - Achievement Causes
   - Achievement Causal Chain
   - Maintenance Causes
   - Actual Cause
3. Relationship with Halpern-Pearl
4. Conclusions

References I

Beckers, S. and Vennekens, J. (2012).
Counterfactual dependency and actual causation in cp-logic and structural models: a comparison.
In *Proceedings of the Sixth Starting AI Researchers' Symposium*, volume 241, pages 35–46.

Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., Spirtes, P., Teng, C. M., and Zhang, J. (2010).
Actual causation: a stone soup essay.
Synthese, 175(2):169–192.

Hall, N. (2000).
Causation and the price of transitivity.
Journal of Philosophy, 97(4):198–222.

Halpern, J. Y. (2015).
A modification of the Halpern-Pearl definition of causality.
In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 3022–3033.

Halpern, J. Y. (2016).
Actual Causality.
The MIT Press, ISBN 9780262035026 edition.

Hitchcock, C. (2007).
Prevention, preemption, and the principle of sufficient reason.
The Philosophical Review, 116(4):495–532.

Hopkins, M. and Pearl, J. (2007).
Causality and counterfactuals in the situation calculus.
Journal of Logic and Computation, 17(5):939–953.

References III

Paul, L. and Hall, N. (2013).
Causation: a user's guide.
Oxford University Press, ISBN 978-0199673452.

Pearl, J. (1998).
On the definition of actual cause.
Technical report, R-259, University of California Los Angeles.