Learning about causation: dynamic epistemic modal logic with causal awareness

Nina Gierasimczuk[1] Hermine Grosinger[2] Katrine B.P. Thoft[1]

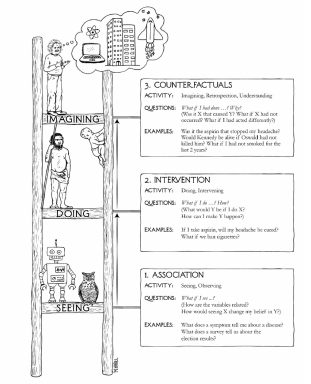
[1]Algorithms, Logic and Graphs
Department of Applied Mathematics and Computer Science
Technical University of Denmark

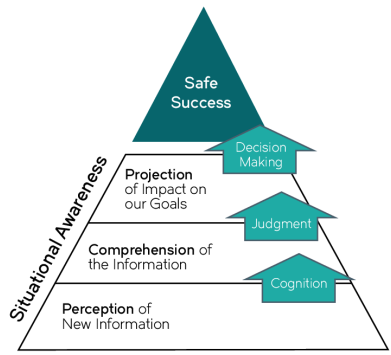
[2]Center for Applied Sensor Systems (AASS)
Örebro University

DaLí Workshop, Xi'An, China
October 20, 2025

Contents

Causality and Situation Awareness





The ladder of causality and the hierarchy of situational awareness.

Pearl, J., Mackenzie, D.: The book of why: the new science of cause and effect. Basic Books (2018)

Krajewski, J.: Situational awareness.The next leap in industrial human machine interface design (2018)

Proactive AI Systems

Proactivity amounts to autonomously initiating action while taking into account future state development as well as anticipating potential consequences of the agent's actions on (the minds of) other agents and the environment.

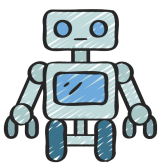

Grosinger, J.: On Proactive Human-AI Systems. AIC 2022


Lorini, E.: Designing artificial reasoners for communication. AAMAS 2024

Proactive AI Systems — Example

Proactive robot:

- *knows* that Bob is
- not *aware* of the medicine;
- wants to *make* him *aware* of the medicine and its
- *causal relations* to his health



In this talk...

- We enrich reasoning about interventions in causal structures with awareness
- We draw from existing lines of research:

  Dynamic epistemic logic of causality with counterfactual reasoning about interventions

  Barbero, F., et al.: Thinking about causation: A causal language with epistemic operators. DaLí 2020

  Modal logic-based dynamic approach to knowledge and awareness

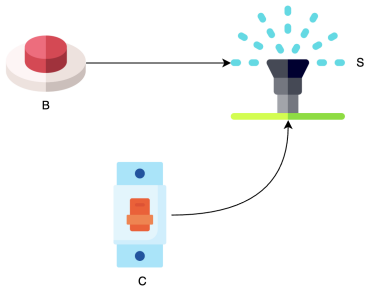
  

  van Benthem, J. and Velázquez-Quesada, F.: The dynamics of awareness. Synthese 2010

- Implicit vs explicit knowledge about causality
- Reasoning about causality on two levels: expert - learner
- Learning about causality and dynamic *awareness-raising*

Contents

Billie's sprinkler

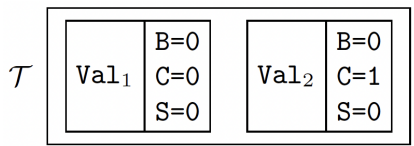


Billie wants to switch on a sprinkler (S), which turns on when she presses the red button (B). However, that will work only if the circuit breaker (C) is closed.

Barbero, F., et al.: Thinking about causation: A causal language with epistemic operators. DaLí 2020.

Billie's sprinkler: Uncertainty

Assume the button is not pressed, the circuit is open and the sprinkler is off, $B = C = S = 0$. Billie knows the values of B and S, but she is uncertain about C.



Epistemic causal models

Definition (Epistemic causal model, Barbero et al. 2022)

An **epistemic causal model** is a triple $E = \langle S, \mathcal{F}, \mathcal{T} \rangle$, where:

- $S = \langle \mathbf{U}, \mathbf{V}, R \rangle$ is a finite with exogenous variables $\mathbf{U} = \{U_1, \ldots, U_m\}$, endogenous variables $\mathbf{V} = \{V_1, \ldots, V_n\}$, and the range $R(X)$ of each variable X;

Epistemic causal models

Definition (Epistemic causal model, Barbero et al. 2022)

An **epistemic causal model** is a triple $E = \langle S, \mathcal{F}, \mathcal{T} \rangle$, where:

- $S = \langle \mathbf{U}, \mathbf{V}, R \rangle$ is a finite with exogenous variables $\mathbf{U} = \{U_1, \ldots, U_m\}$, endogenous variables $\mathbf{V} = \{V_1, \ldots, V_n\}$, and the range $R(X)$ of each variable X;
- $\mathcal{F}$ is the set of structural functions for each $V_j \in \mathbf{V}$, with $f_{V_j} : R(U_1, \ldots, U_m, V_1, \ldots, V_n) \to R(V_j)$.

Epistemic causal models

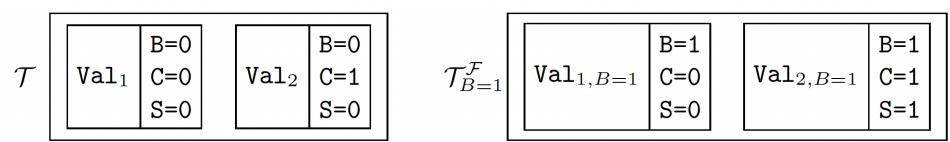
Definition (Epistemic causal model, Barbero et al. 2022)

An **epistemic causal model** is a triple $E = \langle S, \mathcal{F}, \mathcal{T} \rangle$, where:

- $S = \langle \mathbf{U}, \mathbf{V}, R \rangle$ is a finite with exogenous variables $\mathbf{U} = \{U_1, \ldots, U_m\}$, endogenous variables $\mathbf{V} = \{V_1, \ldots, V_n\}$, and the range $R(X)$ of each variable X;
- $\mathcal{F}$ is the set of structural functions for each $V_j \in \mathbf{V}$, with $f_{V_j} : R(U_1, \ldots, U_m, V_1, \ldots, V_n) \to R(V_j)$.
- $\mathcal{T}$ is a set of valuation functions $\texttt{Val}$, where $\texttt{Val}(X) \in R(X)$ for each $X \in \mathbf{U} \cup \mathbf{V}$, s.t. the value of endogenous variable must comply with the structural function of the variable.

Billie's sprinkler: Intervention

What would happened to the uncertainty range if a (counterfactual) intervention is performed? We can perform an intervention of pressing the button, $B{=}1$, and observe the potential change in Billie's knowledge.



Interventions

Definition (Assignment)

An *assignment* on a signature S is an expression $\overrightarrow{X} = \overrightarrow{x}$, where $\overrightarrow{X}$ is a tuple of different variables from $U \cup V$ and $\overrightarrow{x} \in R(\overrightarrow{X})$.

Interventions

Definition (Assignment)

An *assignment* on a signature S is an expression $\overrightarrow{X}=\overrightarrow{x}$, where $\overrightarrow{X}$ is a tuple of different variables from $U \cup V$ and $\overrightarrow{x} \in R(\overrightarrow{X})$.

Definition (Intervention)

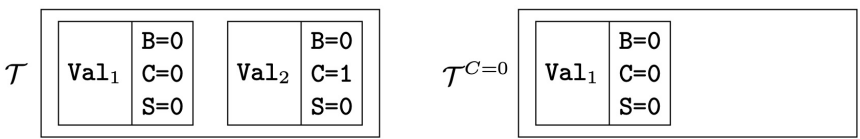
For an epistemic causal model $E = \langle S, \mathcal{F}, \mathcal{T} \rangle$ and an assignment $\overrightarrow{X}=\overrightarrow{x}$, the *intervention* of setting the variables in $\overrightarrow{X}$ to $\overrightarrow{x}$, results in $E_{\overrightarrow{X}=\overrightarrow{x}} = \langle S, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}} \rangle$, where:

- $\mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}$ is the same as $\mathcal{F}$, except that for every endogenous variable X_i in $\overrightarrow{X}$, f_{X_i} is replaced with a constant function returning x_i;
- $\mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}} := \{\mathtt{Val}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}} \mid \mathtt{Val} \in \mathcal{T}\}$, where for each exogenous variable not in $\overrightarrow{X}$, $\mathtt{Val}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}$ is as $\mathtt{Val}$; for each exogenous variable X_i in $\overrightarrow{X}$, $\mathtt{Val}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}(X_i) = x_i$, and the values of all endogenous variables not in $\overrightarrow{X}$ comply with $\mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}$.

Billie's sprinkler: Announcement

Billie has been informed that the circuit-breaker is open, so she can eliminate $\texttt{Val}_2$. As a consequence she would know that had she pushed the button, the sprinkler would be off:

$$[C{=}0!]K[B{=}1]S{=}0$$



The $\mathcal{L}_{\texttt{PAKC}}$ language

$$\gamma ::= Z{=}z \mid \neg\gamma \mid \gamma \wedge \gamma \mid K\gamma \mid [\gamma!]\gamma \qquad \text{for } Z \in U \cup V \text{ and } z \in \mathcal{R}(Z)$$
$$\varphi ::= Z{=}z \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid [\varphi!]\varphi \mid [\overrightarrow{X}{=}\overrightarrow{x}]\gamma \qquad \text{for } \overrightarrow{X}{=}\overrightarrow{x} \text{ on } S$$

The language $\mathcal{L}_{\texttt{PAKC}}$ is interpreted locally, with the formulas evaluated at a pair $(E, \texttt{Val})$, i.e., an epistemic causal model and one of its valuations $\texttt{Val} \in \mathcal{T}$. The Boolean fragment is as usual, and the other clauses have the following meaning:

$$\begin{array}{lll}
(E, \texttt{Val}) \models Z{=}z & \text{iff} & \texttt{Val}(Z) = z \\
(E, \texttt{Val}) \models K\varphi & \text{iff} & (E, \texttt{Val}') \models \varphi \text{ for every } \texttt{Val}' \in \mathcal{T} \\
(E, \texttt{Val}) \models [\psi!]\varphi & \text{iff} & (E, \texttt{Val}) \models \psi \text{ implies } (E^{\psi}, \texttt{Val}) \models \varphi \\
(E, \texttt{Val}) \models [\overrightarrow{X}{=}\overrightarrow{x}]\gamma & \text{iff} & (E_{\overrightarrow{X}=\overrightarrow{x}}, \texttt{Val}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}) \models \gamma
\end{array}$$

where $E^{\psi} = \langle S, \mathcal{F}, \mathcal{T}^{\psi} \rangle$, with $\mathcal{T}^{\psi} := \{\texttt{Val}' \in \mathcal{T} \mid (E, \texttt{Val}') \models \psi\}$.

The 'no learning' effect and logical omniscience

Barbero et al 2020 provide a sound and complete axiomatization of $\mathcal{L}_{\texttt{PAKC}}$.

They note that in their logic the following is a validity:

$$[\overrightarrow{X}=\overrightarrow{x}]K\varphi \to K[\overrightarrow{X}=\overrightarrow{x}]\varphi$$

This law could be seen as a 'no-learning' effect: if after the intervention the agents knows φ, then then the agent knows the intervention will result in φ.

The 'no learning' effect and logical omniscience

Barbero et al 2020 provide a sound and complete axiomatization of $\mathcal{L}_{\texttt{PAKC}}$.

They note that in their logic the following is a validity:

$$[\overrightarrow{X}=\overrightarrow{x}]K\varphi \rightarrow K[\overrightarrow{X}=\overrightarrow{x}]\varphi$$

This law could be seen as a 'no-learning' effect: if after the intervention the agents knows φ, then then the agent knows the intervention will result in φ.

Knowing 'too much' is a known problem in logic. The problem of logical omniscience, i.e., the effect of knowledge being closed on logical consequence, appears by default in epistemic logic and makes the theory incompatible with cognitive reality. While we might know some facts, we might be unaware of some of their consequences.

The 'no learning' effect and logical omniscience

Barbero et al 2020 provide a sound and complete axiomatization of $\mathcal{L}_{\texttt{PAKC}}$.

They note that in their logic the following is a validity:

$$[\overrightarrow{X}=\overrightarrow{x}]K\varphi \rightarrow K[\overrightarrow{X}=\overrightarrow{x}]\varphi$$

This law could be seen as a 'no-learning' effect: if after the intervention the agents knows φ, then then the agent knows the intervention will result in φ.

Knowing 'too much' is a known problem in logic. The problem of logical omniscience, i.e., the effect of knowledge being closed on logical consequence, appears by default in epistemic logic and makes the theory incompatible with cognitive reality. While we might know some facts, we might be unaware of some of their consequences.

Both issues can be addressed by involving awareness.

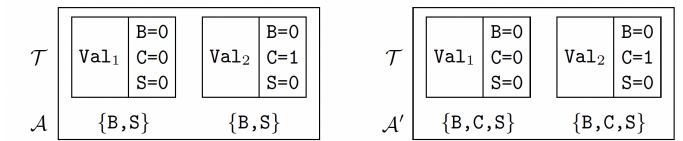
Contents

Billie's sprinker: Causal Awareness

Billie's gardener is an expert on watering the grounds around her mansion. Billie does not have to worry about the circuit breakers— she is aware of the button, and, normally, pressing it turns the sprinkler on. The gardener, on the other hand, is not limited in his awareness of the electrical wiring of the garden, and so of the circuit-breaker. Both the gardener and Billie can directly observe the button and the sprinkler, but they do not see the state of the circuit breaker. As long as the button works as intended, i.e., it turns on the sprinkler, Billie is perfectly happy with her level of awareness of the causal variables. If that fails, however, she would call on the gardener, and he could (by reasoning counterfactually) realise that she should be made aware of the existence of the circuit-breaker (and explain its relation to other variables).

Billie's sprinker: Causal Awareness

Billie's gardener is an expert on watering the grounds around her mansion. Billie does not have to worry about the circuit breakers— she is aware of the button, and, normally, pressing it turns the sprinkler on. The gardener, on the other hand, is not limited in his awareness of the electrical wiring of the garden, and so of the circuit-breaker. Both the gardener and Billie can directly observe the button and the sprinkler, but they do not see the state of the circuit breaker. As long as the button works as intended, i.e., it turns on the sprinkler, Billie is perfectly happy with her level of awareness of the causal variables. If that fails, however, she would call on the gardener, and he could (by reasoning counterfactually) realise that she should be made aware of the existence of the circuit-breaker (and explain its relation to other variables).



Epistemic causal awareness models

Definition (Epistemic causal awareness models)

A *epistemic causal awareness model* is $\mathcal{M} = \langle S, \mathcal{F}, \mathcal{T}, \mathcal{A} \rangle$, where: $S = \langle U, V, R \rangle$ is a signature, $\mathcal{F}$ is a set of structural functions over the set of endogenous variables V, $\mathcal{T}$ is a non-empty set of possible assignments complying with $\mathcal{F}$, and $\mathcal{A}$ is the awareness function $\mathcal{A} : \mathcal{T} \to \mathcal{P}(U \cup V)$.

The $\mathcal{L}_{\texttt{PAKCA+}}$ language

$$\begin{aligned}\gamma &::= Z = z \mid \neg\gamma \mid \gamma \wedge \gamma \mid K\gamma \mid [\gamma!]\gamma \mid A\gamma \mid [+\overrightarrow{X}]\gamma \\ \varphi &::= Z = z \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid [\varphi!]\varphi \mid A\varphi \mid [+\overrightarrow{X}]\varphi \mid [\overrightarrow{X}{=}\overrightarrow{x}]\gamma\end{aligned} \tag{1}$$

$$\begin{aligned}(\mathcal{M}, \texttt{Val}) &\models A\varphi && \text{iff} \quad v(\varphi) \subseteq \mathcal{A}(\texttt{Val}) \\ (\mathcal{M}, \texttt{Val}) &\models [+\overrightarrow{X}]\ \varphi && \text{iff} \quad (\mathcal{M}_{+\overrightarrow{X}}, \texttt{Val}) \models \varphi\end{aligned}$$

The $\mathcal{L}_{\texttt{PAKCA+}}$ language, cntd.

$(\mathcal{M}, \texttt{Val}) \models A\varphi$ iff $v(\varphi) \subseteq \mathcal{A}(\texttt{Val})$

$v(\varphi)$ is defined in the following way:

$$
\begin{aligned}
&v(Z = z) := \{Z\} \\
&v(\neg\varphi) := v(\varphi) \\
&v([\varphi_1!]\varphi_2) := v(\varphi_1) \cup v(\varphi_2) \\
&v(\varphi_1 \wedge \varphi_2) := v(\varphi_1) \cup v(\varphi_2) \\
&v(\bigcirc\varphi) := v(\varphi), \text{ where } \bigcirc \in \{A, K\} \\
&v(\oplus\varphi) := set(\overrightarrow{X}) \cup v(\varphi), \text{ where} \\
&\oplus \in \{[\overrightarrow{X}{=}\overrightarrow{x}], [+\overrightarrow{X}]\}
\end{aligned}
$$

$(\mathcal{M}, \texttt{Val}) \models [+\overrightarrow{X}]\ \varphi$ iff $(\mathcal{M}_{+\overrightarrow{X}}, \texttt{Val}) \models \varphi$

$\mathcal{M}_{+\overrightarrow{X}}$ is defined in the following way:

Given an epistemic causal awareness model $\mathcal{M} = \langle S, \mathcal{F}, \mathcal{T}, \mathcal{A}\rangle$ and a set of variables $X \subseteq U \cup V$, $\mathcal{M}_{+\overrightarrow{X}} = \langle S, \mathcal{F}, \mathcal{T}, \mathcal{A}'\rangle$, where $\mathcal{A}'(\texttt{Val}) = \mathcal{A}(\texttt{Val}) \cup set(\overrightarrow{X})$, for all $\texttt{Val} \in \mathcal{T}$.

van Ditmarsch, et al.: Implicit, explicit and speculative knowledge. Artificial Intelligence 2018.

van Benthem, J. and Velázquez-Quesada, F.R.: The dynamics of awareness. Synthese 2010.

Explicit causal knowledge

Explicit knowledge is defined in the following way (like in vB+VQ 2010):

$$K^{Ex}\varphi := K(\varphi \wedge A\varphi).$$

in contrast to Halpern's definition: $K\varphi \wedge A\varphi$.

Proposition

If $\models A\varphi \rightarrow KA\varphi$, *then* $\models (K\varphi \wedge A\varphi) \leftrightarrow K(\varphi \wedge A\varphi)$.

As it turns out, weak introspection $\models A\varphi \rightarrow KA\varphi$ implies *uniform awareness*.

Proposition

Let $\mathcal{M} = \langle S, \mathcal{F}, \mathcal{T}, \mathcal{A} \rangle$. $\mathcal{M} \models A\varphi \rightarrow KA\varphi$ *if and only if for all* $\mathtt{Val}, \mathtt{Val}' \in \mathcal{T}$, $\mathcal{A}(\mathtt{Val}) = \mathcal{A}(\mathtt{Val}')$.

Explicit causal intervention and... 'yes-learning'

$$\models [\overrightarrow{X}{=}\overrightarrow{x}]K\varphi \leftrightarrow K[\overrightarrow{X}{=}\overrightarrow{x}]\varphi$$

$$\models [\overrightarrow{X}{=}\overrightarrow{x}]K^{Ex}\varphi \leftrightarrow K^{Ex}[\overrightarrow{X}{=}\overrightarrow{x}]\varphi$$

Explicit causal intervention and... 'yes-learning'

$$\models [\overrightarrow{X}{=}\overrightarrow{x}]K\varphi \leftrightarrow K[\overrightarrow{X}{=}\overrightarrow{x}]\varphi$$

$$\models [\overrightarrow{X}{=}\overrightarrow{x}]K^{Ex}\varphi \leftrightarrow K^{Ex}[\overrightarrow{X}{=}\overrightarrow{x}]\varphi$$

We define explicit causal intervention:

$$[\overrightarrow{X}{=}\overrightarrow{x}]^{Ex} := [\overrightarrow{X}{=}\overrightarrow{x}][+\overrightarrow{X}]$$

aka, expert uttering the counterfactual reasoning in the presence of the layperson.

The order of the two operations does not matter: $\models [\overrightarrow{X}{=}\overrightarrow{x}][+\overrightarrow{X}]\varphi \leftrightarrow [+\overrightarrow{X}][\overrightarrow{X}{=}\overrightarrow{x}]\varphi$, because they change different components of the model.

Explicit causal intervention and... 'yes-learning'

$$\models [\overrightarrow{X}{=}\overrightarrow{x}]K\varphi \leftrightarrow K[\overrightarrow{X}{=}\overrightarrow{x}]\varphi$$

$$\models [\overrightarrow{X}{=}\overrightarrow{x}]K^{Ex}\varphi \leftrightarrow K^{Ex}[\overrightarrow{X}{=}\overrightarrow{x}]\varphi$$

We define explicit causal intervention:

$$[\overrightarrow{X}{=}\overrightarrow{x}]^{Ex} := [\overrightarrow{X}{=}\overrightarrow{x}][+\overrightarrow{X}]$$

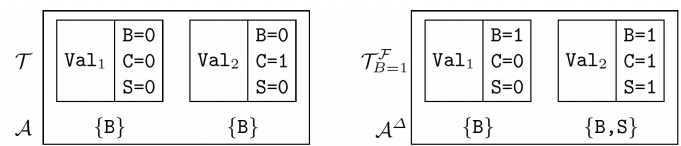
aka, expert uttering the counterfactual reasoning in the presence of the layperson.

The order of the two operations does not matter: $\models [\overrightarrow{X}{=}\overrightarrow{x}][+\overrightarrow{X}]\varphi \leftrightarrow [+\overrightarrow{X}][\overrightarrow{X}{=}\overrightarrow{x}]\varphi$, because they change different components of the model.

$$\not\models [\overrightarrow{X}{=}\overrightarrow{x}]^{Ex}K^{Ex}\varphi \rightarrow K^{Ex}[\overrightarrow{X}{=}\overrightarrow{x}]^{Ex}\varphi$$

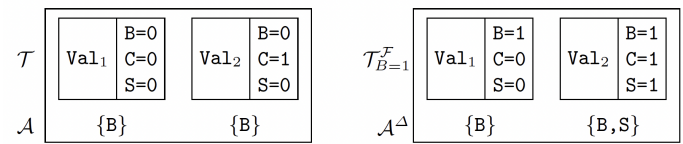
Billie's sprinkler: Differential Awareness

Let us now return to our Billie example, and this time let us assume that Billie is only aware of the button. The expert can counterfactually reason that had the button been pressed, the sprinkler would turn on, and would be noticed by Billie (she would become aware of the variable S). That would happen only in the possible world in which the circuit-breaker is closed.



Billie's sprinkler: Differential Awareness

Let us now return to our Billie example, and this time let us assume that Billie is only aware of the button. The expert can counterfactually reason that had the button been pressed, the sprinkler would turn on, and would be noticed by Billie (she would become aware of the variable S). That would happen only in the possible world in which the circuit-breaker is closed.



Note that this requires non-uniform awareness, so no weak introspection.

Differential causal awareness

- New language $\mathcal{L}_{\texttt{PAKCA}+\Delta}$
- New type of explicit intervention $[\overrightarrow{X}{=}\overrightarrow{x}]^{\Delta}$
- only those variables that change are added to the awareness set.
- Semantics:

$$(\mathcal{M}, \texttt{Val}) \models [\overrightarrow{X}{=}\overrightarrow{x}]^{\Delta}\varphi \text{ iff } (\mathcal{M}_{+\Delta}, \texttt{Val}) \models \varphi,$$

where:
$\mathcal{M}_{+\Delta} = \langle S, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\Delta}\rangle$,
for any $\texttt{Val} \in \mathcal{T}$, $\mathcal{A}^{\Delta}(\texttt{Val}) = \mathcal{A}(\texttt{Val}) \cup \Delta(\texttt{Val}, \texttt{Val}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}})$, and
$\Delta(\texttt{Val}, \texttt{Val}') = \{X \in U \cup V \mid \texttt{Val}(X) \neq \texttt{Val}'(X)\}$

Contents

Hybrid Causal Graphs for Proactivity

Imagine Billie is aware of the button (B) and sprinkler (S).

Hybrid Causal Graphs for Proactivity

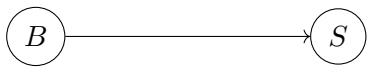
Imagine Billie is aware of the button (B) and sprinkler (S).

The gardener has the full picture, that is, additionally he is aware of:
the circuit breaker (C) and the tap (T), and he knows their causal dependencies.

How can the gardener proactively decide which part of the actual causal structure to reveal to Billie?

Billie's sprinkler

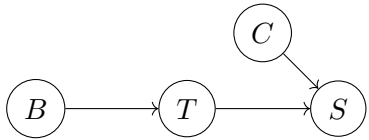
Billie's sprinkler: Learner



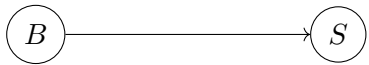
Learner

Billie's sprinkler: Learner and Expert

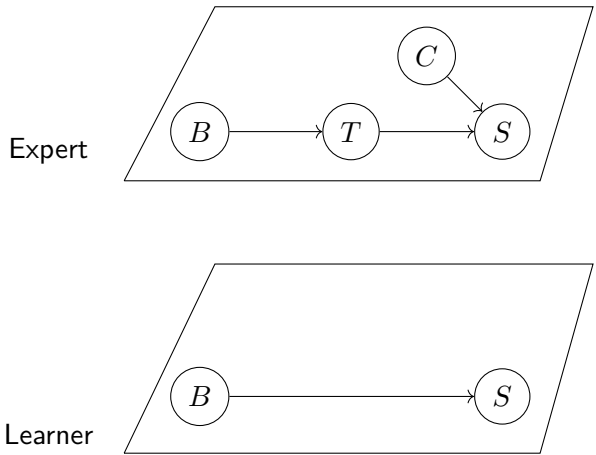
Expert



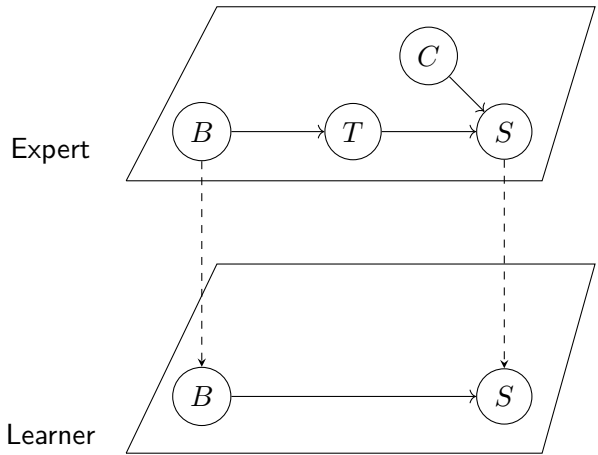
Learner



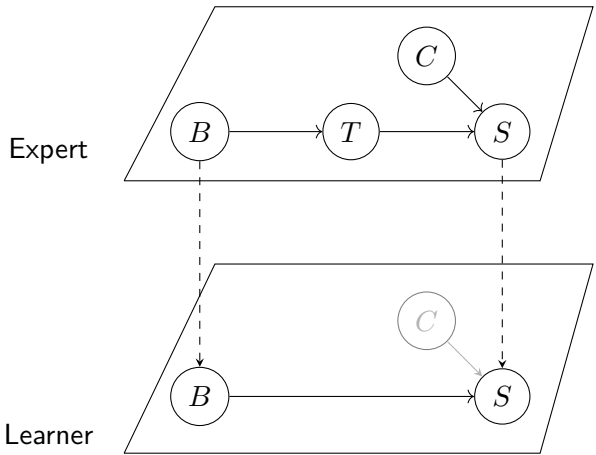
Billie's sprinkler: Learner and Expert



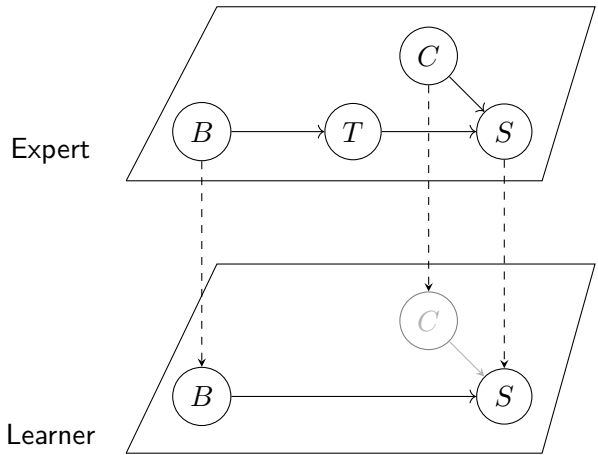
Billie's sprinkler: Learner and Expert



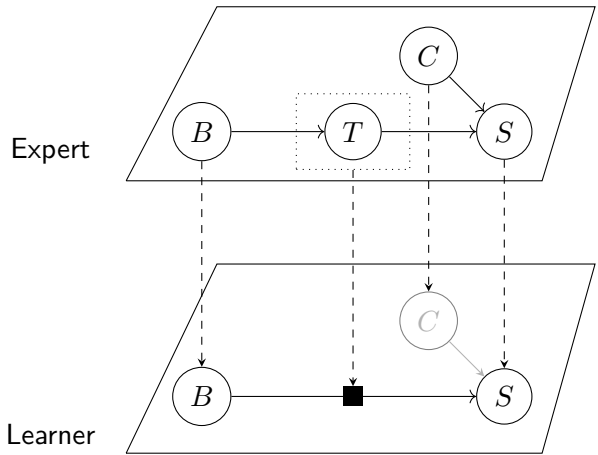
Billie's sprinkler: Learner and Expert



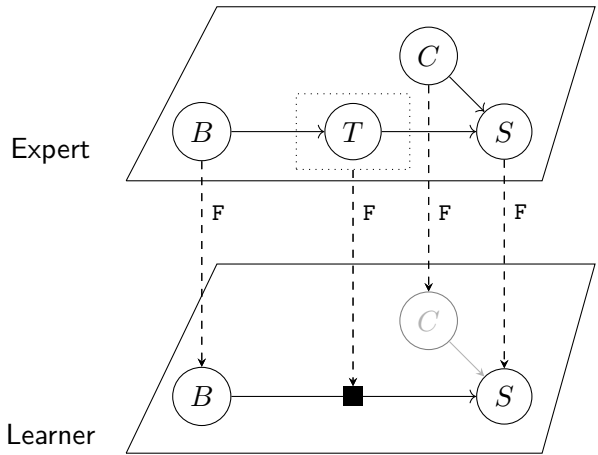
Billie's sprinkler: Learner and Expert



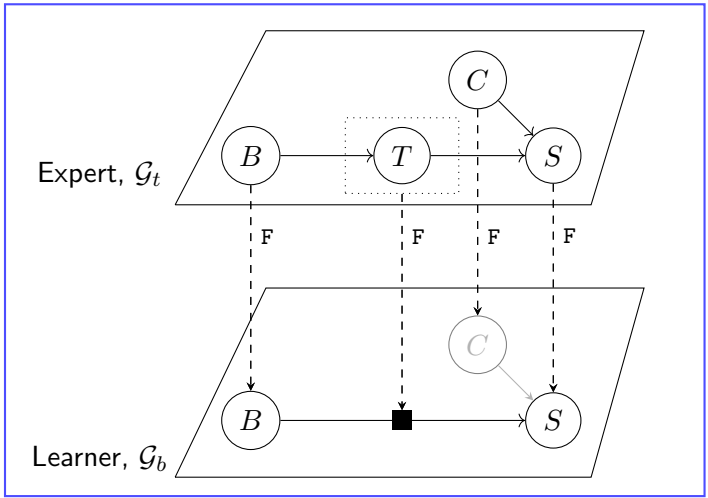
Billie's sprinkler: Learner and Expert



Billie's sprinkler: Learner and Expert



Hybrid Causal Graph



$\mathcal{G} = (\mathcal{G}_t, \mathcal{G}_b, \mathtt{F})$

Contents

Conclusion

- A new **dynamic epistemic logic** with public announcements, **causal interventions** and **dynamic awareness**, $\mathcal{L}_{\texttt{PAKCA+}}$ (and $\mathcal{L}_{\texttt{PAKCA+}\Delta}$)

Conclusion

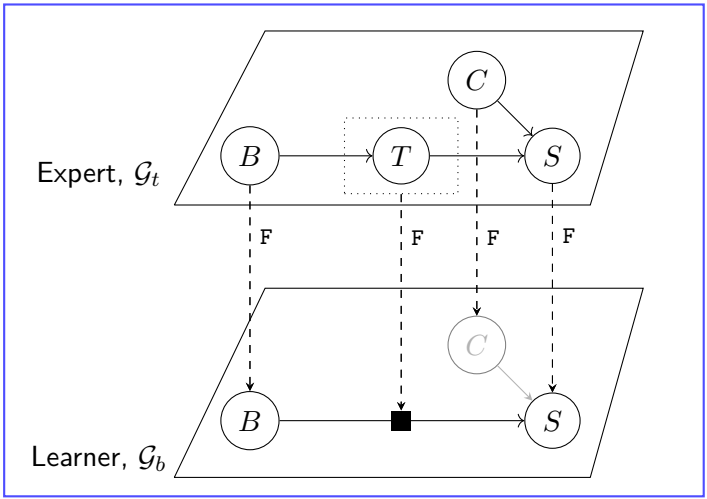
- A new **dynamic epistemic logic** with public announcements, **causal interventions** and **dynamic awareness**, $\mathcal{L}_{\texttt{PAKCA+}}$ (and $\mathcal{L}_{\texttt{PAKCA+}\Delta}$)

- We introduce a new notion of **explicit interventions** with **differential awareness**, and show that the inclusion of explicit knowledge and explicit awareness invalidates the (adjusted) principle of *no-learning*

Conclusion

- A new **dynamic epistemic logic** with public announcements, **causal interventions** and **dynamic awareness**, $\mathcal{L}_{\texttt{PAKCA+}}$ (and $\mathcal{L}_{\texttt{PAKCA+}\Delta}$)
- We introduce a new notion of **explicit interventions** with **differential awareness**, and show that the inclusion of explicit knowledge and explicit awareness invalidates the (adjusted) principle of *no-learning*
- We introduce *hybrid causal graph* for an expert who has **implicit knowledge** as a guideline to make **proactive** decisions about which hidden causal variables to reveal to the layperson who has **explicit knowledge**.

Discussion: Proactive Human-AI Systems



$\mathcal{G} = (\mathcal{G}_t, \mathcal{G}_b, \mathtt{F})$

Discussion: Proactive Human-AI system

Proactive Human-AI Systems can use our modeling for reasoning with causality, awareness, and knowledge

Discussion: Proactive Human-AI system

Proactive Human-AI Systems can use our modeling for reasoning with causality, awareness, and knowledge

Billie's sprinkler example

What should Billie be made aware of
to achieve short-term goals (sprinkler is on) or maintenance goals (grass stays green)?

Discussion: Proactive Human-AI system

Proactive Human-AI Systems can use our modeling for reasoning with causality, awareness, and knowledge

Billie's sprinkler example

What should Billie be made aware of
to achieve short-term goals (sprinkler is on) or maintenance goals (grass stays green)?

If the 'AI-gardener' knows about Billie's *awareness* of the button and the sprinkler and her *unawareness* of the circuit, then it can proactively inform her about the circuit and its *causal link* to the sprinkler when it anticipates that Billie aims to turn on the sprinkler.

Future work: Multiple agents

Support multiple learners and multiple experts — several layers in the Hybrid Causal graph

Future work: Speculative knowledge

Support multiple learners and multiple experts — several layers in the Hybrid Causal graph

Investigate more the meaning *shadow variables* and extend our framework with *speculative knowledge* which concerns anything an agent believes might be possible based on their awareness. The learner could perform counterfactual interventions on shadow variables (learner speculatively knows).

Future work

Support multiple learners and multiple experts — several layers in the Hybrid Causal graph

Investigate more the meaning *shadow variables* and extend our framework with *speculative knowledge* which concerns anything an agent believes might be possible based on their awareness. The learner could perform counterfactual interventions on shadow variables (learner speculatively knows).

Introduce a becoming unaware operator, that is, *drop*-operator $[-\chi]$

Do a sound and complete axiomatization of our logic, $\mathcal{L}_{PAKCA+}$ (and $\mathcal{L}_{PAKCA+\Delta}$).

Thank you!

nigi@dtu.dk
hermine.grosinger@oru.se
kabpt@dtu.dk

References


Barbero, F., Schulz, K., Smets, S., Velazquez-Quesada, F. : Thinking About Causation: A Causal Language with Epistemic Operators. In: Dynamic Logic. New Trends and Applications in Proceedings, pp. 17–32. Springer International Publishing (2020)


van Benthem, J., Velazquez-Quesada, F.: The Dynamics of Awareness. Synthese **177**(12), pp. 5–27, 2010


van Ditmarch, H., French, T., Velázquez-Quesada, F., Wáng, Y.: Implicit, explicit and speculative knowledge. Artificial Intelligence, 35–67 (2017)


Grosinger, J. On Proactive Human-AI Systems. AIC (pp. 140-146). 2022.


Halpern, J., Piermont, E. .: Dynamic Awareness. CoRR (2020)


Lorini, E. Designing Artificial Reasoners for Communication. Proc. AAMAS, pp. 2690-2695. 2024.


Pearl, J., Mackenzie, D.: The book of why: the new science of cause and effect. Basic Books (2018)


Thoft, K., Gierasimczuk, N.: Learning by Intervention in Simple Causal Domains. Lecture Notes in Computer Science, pp. 104–118, 2024

Free variables

Definition

Let φ be a formula of $\mathcal{L}_{\texttt{PAKCA+}}$. We define inductively the set of variables of a φ, $v(\varphi)$, in the following way:

$$
\begin{aligned}
&v(Z = z) := \{Z\}\\
&v(\neg\varphi) := v(\varphi)\\
&v([\varphi_1!]\varphi_2) := v(\varphi_1) \cup v(\varphi_2)\\
&v(\varphi_1 \wedge \varphi_2) := v(\varphi_1) \cup v(\varphi_2)\\
&v(\bigcirc\varphi) := v(\varphi), \text{ where } \bigcirc \in \{A, K\}\\
&v(\oplus\varphi) := set(\overrightarrow{X}) \cup v(\varphi), \text{ where}\\
&\oplus \in \{[\overrightarrow{X}{=}\overrightarrow{x}], [+\overrightarrow{X}]\}
\end{aligned}
$$

Hybrid causal graph definition

Definition

Let $\mathcal{M} = \langle S, \mathcal{F}, \mathcal{T}, \mathcal{A} \rangle$ be an epistemic causal awareness model with uniform awareness.

1. The *implicit causal graph* of $\mathcal{M}$ is $\mathcal{G}_{\mathcal{M}} = (\mathcal{X}, \mathcal{E})$, where $\mathcal{X} = U \cup V$, and $\mathcal{E} = \{(X, Y) \in (U \cup V)^2 \mid X \hookrightarrow_{\mathcal{F}} Y\}$.
2. The *explicit causal graph* of $\mathcal{M}$ is the implicit causal graph of $\mathcal{M}$ restricted to the awareness set, i.e., $\mathcal{G}^{Ex}_{\mathcal{M}} = (\mathcal{X}^{Ex}, \mathcal{E}^{Ex})$, where $\mathcal{X}^{Ex} = \mathbf{A}_{\mathcal{M}}$, and $\mathcal{E}^{Ex} = \{(X, Y) \in (\mathbf{A}_{\mathcal{M}})^2 \mid X \hookrightarrow^{+}_{\mathcal{F}} Y\}$.
3. Let $\mathcal{G}_{\mathcal{M}} = (\mathcal{X}, \mathcal{E})$ be the implicit causal graph of $\mathcal{M}$. The *one-step closure* of the explicit causal graph $\mathcal{G}^{Ex}_{\mathcal{M}} = (\mathcal{X}^{Ex}, \mathcal{E}^{Ex})$ is $\hat{\mathcal{G}}^{Ex}_{\mathcal{M}} = (\hat{\mathcal{X}}^{Ex}, \hat{\mathcal{E}}^{Ex})$, where:
   - $\hat{\mathcal{X}}^{Ex} = \mathcal{X}^{Ex} \cup \{Y \in \mathcal{X} \setminus \mathbf{A}_{\mathcal{M}} \mid (X, Y) \in \mathcal{E} \text{ or } (Y, X) \in \mathcal{E} \text{ for some } X \in \mathcal{X}^{Ex}\}$;
   - $\hat{\mathcal{E}}^{Ex} = \mathcal{E}^{Ex} \cup \{(X, Y) \mid X \in \mathcal{X}^{Ex}, Y \in \hat{\mathcal{X}}^{Ex} \text{ and } (X, Y) \in \mathcal{E}\}$.
4. An edge $(X, Y) \in \mathcal{E}^{Ex}$ is a *black box* if $X \hookrightarrow^{+}_{\mathcal{F}} Y$, but it is not the case that $X \hookrightarrow_{\mathcal{F}} Y$. We denote by $\blacksquare(\mathcal{G}^{Ex}_{\mathcal{M}})$ the set of all such black boxes in $\mathcal{G}^{Ex}_{\mathcal{M}}$.

Defining the hybrid causal graphs

Definition (Hybrid causal graph)

Let $\mathcal{M} = \langle S, \mathcal{F}, \mathcal{T}, \mathcal{A} \rangle$ be an epistemic causal awareness model with uniform awareness. A *hybrid causal graph* of $\mathcal{M}$ is a tuple $\mathcal{G}_{\mathbb{C}} = (\mathcal{G}_t, \mathcal{G}_b, \mathtt{F})$, such that: $\mathcal{G}_t := \mathcal{G}_{\mathcal{M}}$, $\mathcal{G}_b := \hat{\mathcal{G}}^{Ex}_{\mathcal{M}}$, and $\mathtt{F} := \{\mathtt{F}_{sh}, \mathtt{F}_{\blacksquare}\}$, where:

- $\mathtt{F}_{sh} : \mathcal{X}_t \to \hat{\mathcal{X}}_b$ is a partial identity function, with $\mathtt{F}_{sh}(X) = X$, if $X \in \hat{\mathcal{X}}_b$;
- $\mathtt{F}_{\blacksquare} : \mathcal{P}(\mathcal{X}_t) \to \blacksquare(\mathcal{G}_b)$ is a partial function, with $\mathtt{F}_{\blacksquare}(\mathbf{Z}) = (X, Y)$, if $\mathbf{Z} = \{Z \mid X \hookrightarrow^{+}_{\mathcal{F}} Z \hookrightarrow^{+}_{\mathcal{F}} Y\}$.

The consider algorithm

Algorithm 3 CONSIDER$(\mathcal{M}, \overrightarrow{X})$

Input $\overrightarrow{X}$: a sequence of variables

$\mathcal{M} = \langle S, \mathcal{F}, \mathcal{T}, \mathcal{A} \rangle$: an epistemic causal awareness model

Output $\mathcal{M}'$: an updated model after awareness expansion

1: $\mathcal{A}' = null$
2: **for each** $\texttt{Val} \in \mathcal{T}$ **do**
3: $\quad \mathcal{A}'(\texttt{Val}) = \mathcal{A}(\texttt{Val}) \cup set(\overrightarrow{X})$
4: **end for**
5: $\mathcal{M}' = \langle S, \mathcal{F}, \mathcal{T}, \mathcal{A}' \rangle$
6: **return** $\mathcal{M}'$

The intervene algorithm

Algorithm 4 Intervene($\mathcal{M}, \overrightarrow{X}{=}\overrightarrow{x}$)

Input $\mathcal{M} = \langle\langle U, V, \mathcal{R}\rangle, \mathcal{F}, \mathcal{T}, \mathcal{A}\rangle$: an epistemic causal awareness model, where $\mathcal{F} = \{f_{X_i} \mid X_i \in V\}$ and $\mathcal{T} = \{\texttt{Val} \mid \texttt{Val}$ complies with $\mathcal{F}\}$
$\overrightarrow{X}{=}\overrightarrow{x}$: an assignment
Output $\mathcal{M}'$: the updated model after intervention

$\mathcal{F}' = \mathcal{F}$, $\mathcal{T}' = \mathcal{T}$
for each X_i in $\overrightarrow{X}$ **do**
  if $X_i \in V$ **then**
    f'_{X_i} = constant function x_i
    for each $\texttt{Val}' \in \mathcal{T}'$ **do**
      $\texttt{Val}'(X_i) = x_i$
    end for
  else
    for each $\texttt{Val}' \in \mathcal{T}'$ **do**
      $\texttt{Val}'(X_i) = x_i$
    end for
  end if
end for
for each $X_j \in V \setminus set(\overrightarrow{X})$ **do**
  for each $\texttt{Val}' \in \mathcal{T}'$ **do**
    $\texttt{Val}'(X_j)$ = value complying with $\mathcal{F}'$
  end for
end for
$\mathcal{M}' = \langle\langle U, V, \mathcal{R}\rangle, \mathcal{F}', \mathcal{T}', \mathcal{A}\rangle$
return $\mathcal{M}'$

The explicit intervene algorithm

Algorithm 1 InterveneExplicit($\mathcal{M}, \overrightarrow{X}=\overrightarrow{x}$)

Input
$\mathcal{M} = \langle S, \mathcal{F}, \mathcal{T}, \mathcal{A} \rangle$: an epistemic causal awareness model
$\overrightarrow{X}=\overrightarrow{x}$: an assignment
Output $\mathcal{M}'$: the updated model after explicit intervention

1: $\mathcal{M}_+ \leftarrow$ Consider($\mathcal{M}, \overrightarrow{X}$)
2: $\mathcal{M}' \leftarrow$ Intervene($\mathcal{M}_+, \overrightarrow{X}=\overrightarrow{x}$)
3: **return** $\mathcal{M}'$

The differential awareness intervene algorithm

Algorithm 2 INTERVENEDIFFAWARE($\mathcal{M}, \overrightarrow{X}=\overrightarrow{x}$)

Input

$\mathcal{M} = \langle S, \mathcal{F}, \mathcal{T}, \mathcal{A} \rangle$: an epistemic causal awareness model

$\overrightarrow{X}=\overrightarrow{x}$: an assignment

Output

$\mathcal{M}^{\Delta}$: the updated model after differential awareness intervention

$\mathcal{A}' = null$

$\mathcal{M}_{\overrightarrow{X}=\overrightarrow{x}} = \langle S, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A} \rangle \leftarrow$ INTERVENE($\mathcal{M}, \overrightarrow{X}=\overrightarrow{x}$)

for all $\texttt{Val} \in \mathcal{T}$ **do**

    $\mathcal{A}'(\texttt{Val}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}) = \mathcal{A}(\texttt{Val}) \cup \Delta(\texttt{Val}, \texttt{Val}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}})$

end for

$\mathcal{M}^{\Delta} = (S, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}')$

return $\mathcal{M}^{\Delta}$