Non-Classical Logics for Explanations in AI Systems

Emiliano Lorini
IRIT-CNRS, Toulouse University, France

9 November 2022
XAISeminar@Imperial

Logic-based approaches to explanation

- Different notions of explanation are studied in the XAI domain
  - Abductive
  - Contrastive
  - Counterfactual
- Logic-based XAI mostly based on propositional logic (PL)

Logic-based approaches to explanation

- Fundamental building blocks of explanation:
  - Counterfactual dependence
  - Variance/invariance:

    > *[I]nvariance is a modal notion – it has to do with whether a relationship would remain stable under various hypothetical changes [Woodward 2002, p. 225].*

  - Imperfect knowledge of the classifier ("black box")
    $\Rightarrow$ Epistemic/subjective explanation
- Beyond PL: need for more expressive languages
- Non-classical logics:
  - Modal logic (ML)
  - Conditional logic (CL)
  - Epistemic logic (EL), dynamic EL (DEL)
  - Deontic logic (DL)

Outline

1. Explanations in “white box” classifiers
2. Explanations in “black box” classifiers
3. Open problems and future extensions

Outline

1. Explanations in “white box” classifiers
2. Explanations in “black box” classifiers
3. Open problems and future extensions

Reasoning about "white box" classifiers

Main idea: a binary input classifier is a partition of all possible input instances in an S5 Kripke model

Permanent job (*pe*)	> 3000 € monthly salary (*sa*)	EU citizenship (*eu*)	Loan
0	0	0	**No**
0	0	1	**No**
0	1	0	**No**
1	0	0	**Yes**
0	1	1	**Yes**
1	0	1	**Yes**
1	1	0	**Yes**
1	1	1	**Yes**

Figure: A classifier

States/instances	f_1
$s_1=\{\}$	No
$s_2=\{eu\}$	No
$s_3=\{sa\}$	No
$s_4=\{pe\}$	Yes
$s_5=\{sa,eu\}$	Yes
$s_6=\{pe,eu\}$	Yes
$s_7=\{pe,sa\}$	Yes
$s_8=\{pe,sa,eu\}$	Yes

Figure: Its S5 representation

Formal semantics

- Atm_0: countable set of atoms representing input features
- Val: finite set of classification values (or classes)

Definition (Classifier model)

A classifier model (CM) is a tuple $C = (S, f)$ where

- $S \subseteq 2^{Atm_0}$ is a set of input instances,
- $f : S \longrightarrow Val$ is a classification function.

Modal language

$$\varphi \quad ::= \quad p \mid \mathsf{t}(x) \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_{\mathtt{I}}\varphi$$

with p ranging over Atm_0 and x ranging over Val

$\mathsf{t}(x) \approx$ "the actual input instance is classified as x"

$\Box_{\mathtt{I}}\varphi \approx$ "the classifier *necessarily* satisfies φ"

$\approx$ "φ is true for all input instances of the classifier"

Semantic interpretation wrt CM $C = (S, f)$ and $s \in S$ (pointed CM):

$$\begin{aligned} (C,s) \models p &\iff p \in s \\ (C,s) \models \mathsf{t}(x) &\iff f(s) = x \\ (C,s) \models \Box_{\mathtt{I}}\varphi &\iff \forall s' \in S : (C,s') \models \varphi \end{aligned}$$

Useful "ceteris paribus" modalities

Let $X \subseteq Atm_0$ finite:

$$[X]\varphi =_{def} \bigwedge_{Y \subseteq X} \Big(\big(\bigwedge_{p \in Y} \wedge \bigwedge_{p \in X \setminus Y} \neg p \big) \rightarrow \Box_{\mathtt{I}} \big(\big(\bigwedge_{p \in Y} \wedge \bigwedge_{p \in X \setminus Y} \neg p \big) \rightarrow \varphi \big) \Big)$$

We have:

$$(C, s) \models [X]\varphi \iff \forall s' \in S : \text{ if } (s \cap X) = (s' \cap X) \text{ then } (C, s') \models \varphi$$

$[X]\varphi \approx$ "φ is true all atoms in X being equal"

$\approx$ "φ is true regardless of the value of the atoms in $Atm_0 \setminus X$"

Useful “ceteris paribus” modalities

- Connection with prop. dependence logic [Yang & Väänänen, 2016]
- Dependence atom (“q only depends on $p_1, \dots, p_k$”):

$$\begin{aligned}\mathrm{Dep}(p_1, \dots, p_k, q) =_{def} & [\emptyset]\big(q \to [\{p_1, \dots, p_k\}]q\big) \wedge \\ & [\emptyset]\big(\neg q \to [\{p_1, \dots, p_k\}]\neg q\big)\end{aligned}$$

Complexities

	Finite (fixed) variables	**Infinite variables**
Modalities $[X]$ **are defined as abbreviations**	Polynomial	NP-complete
Modalities $[X]$ **are primitives**	Polynomial	NEXPTIME-complete

Table: Summary of complexity results

Explanations

Let λ be a term (conjunction of literals):

- Prime implicant:

$$\mathtt{PImp}(\lambda, x) =_{def} [\emptyset]\Big(\lambda \to \big(\mathsf{t}(x) \wedge \bigwedge_{p \in Atm(\lambda)} \langle Atm(\lambda) \setminus \{p\}\rangle \neg \mathsf{t}(x)\big)\Big)$$

- Abductive explanation:

$$\mathtt{AXp}(\lambda, x) =_{def} \lambda \wedge \mathtt{PImp}(\lambda, x)$$

- Contrastive explanation:

$$\mathtt{CXp}(\lambda, x) =_{def} \lambda \wedge \langle Atm_0 \setminus Atm(\lambda)\rangle \neg \mathsf{t}(x) \wedge \bigwedge_{p \in Atm(\lambda)} [(Atm_0 \setminus Atm(\lambda)) \cup \{p\}]\mathsf{t}(x)$$

Explanations

States/instances	f_1
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	No
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	Yes
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

$s_4 \models \texttt{AXp}(pe, Yes)$

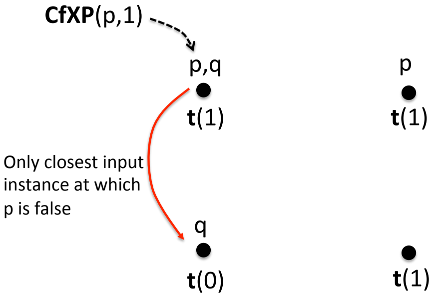
$s_2 \models \texttt{CXp}(\neg sa, No)$

Explanations

- <u>Counterfactual explanation</u>:

$$\color{red}{\texttt{CfXp}(\varphi, x)} =_{def} \mathrm{t}(x) \wedge \left(\neg\varphi \Rightarrow \neg\mathrm{t}(x)\right)$$

Remark: Lewis-like conditional $\Rightarrow$ defined as an abbreviation in finite-variable case (semantics based on Hamming dist.)



Explanations

States/instances	f_1
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	No
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	Yes
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

$s_2 \models \texttt{CfXp}(\neg sa, No)$

Some remarkable properties

⇒ Principle of sufficient reason (PSR):

> *Of everything whatsoever a cause or reason must be assigned, either for its existence, or for its non-existence. [Spinoza, Ethics, 1p11d2]*

$$\models_{\textbf{Definite}} \mathtt{t}(x) \rightarrow \bigvee_{\lambda \in \mathit{Term}} \mathtt{AXp}(\lambda, x)$$

⇒ 'Atomic' CfXp and CXp coincide

$$\models \mathtt{CXp}(l, x) \leftrightarrow \mathtt{CfXp}(l, x) \text{ with } l \text{ a literal}$$

Bias

Local bias:

$$\mathtt{Bias}(x) =_{def} \mathtt{t}(x) \wedge \langle \mathsf{NF} \rangle \neg \mathtt{t}(x)$$

with PF the set of protected features and $\mathsf{NF} = Atm_0 \setminus \mathsf{PF}$

$$\models \mathtt{Bias}(x) \leftrightarrow \bigvee_{Atm(\lambda) \subseteq \mathsf{PF}} \mathtt{CXp}(\lambda, x)$$

Global bias:

$$\mathtt{GBias} =_{def} \Diamond_{\mathtt{I}} \Big(\bigvee_{x \in Val} \mathtt{Bias}(x) \Big)$$

Bias

Let $\text{PF} = \{eu\}$

States/instances	f_1
s_1={}	No
s_2={eu}	No
s_3={sa}	No
s_4={pe}	Yes
s_5={sa,eu}	Yes
s_6={pe,eu}	Yes
s_7={pe,sa}	Yes
s_8={pe,sa,eu}	Yes

$s_3 \models \texttt{Bias}(No)$

Outline

From "white box" to "black box" classifiers

- Two-dimensional semantics: instance×classifier
- Horizontal dimension ≈ uncertainty about classifier's properties
- Bimodal language

	f_1
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	No
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	Yes
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

	f_2
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	No
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	No
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

Bimodal language

$$\varphi \quad ::= \quad p \mid \mathrm{t}(x) \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_{\mathtt{I}}\varphi \mid \Box_{\mathtt{F}}\varphi$$

with p ranging over Atm_0 and x ranging over Val

$\Box_{\mathtt{I}}\varphi \approx$ "the actual classifier *necessarily* satisfies φ,
(regardless of the input instance)"

$\Box_{\mathtt{F}}\varphi \approx$ "the actual input instance *necessarily* satisfies φ,
(regardless of the classifier)"

Two-dimensional semantics

Definition

A multi-classifier model (MCM) is a pair $\Gamma = (S, \Phi)$ where:

- $S \subseteq 2^{Atm_0}$ (set of input instances),
- $\Phi \subseteq Val^S$ (set of possible classifiers).

Semantic interpretation of formulas wrt pointed MCM (Γ, s, f) with $\Gamma = (S, \Phi)$ an MCM, $s \in S$ and $f \in \Phi$:

$$
\begin{aligned}
(\Gamma, s, f) \models p &\iff p \in s \\
(\Gamma, s, f) \models \mathsf{t}(x) &\iff f(s) = x \\
(\Gamma, s, f) \models \Box_{\mathtt{I}}\varphi &\iff \forall s' \in S : (\Gamma, s', f) \models \varphi \\
(\Gamma, s, f) \models \Box_{\mathtt{F}}\varphi &\iff \forall f' \in \Phi : (\Gamma, s, f') \models \varphi
\end{aligned}
$$

Two-dimensional semantics

Bob

	f_1
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	No
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	Yes
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

	f_2
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	No
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	No
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

Bob is $\{sa\}$ and (only) knows that:

- his application was unsuccessful
- necessarily not having a permanent job and not having a good salary will make loan application unsuccessful
- necessarily having a permanent job will make loan application successful

$$
\begin{aligned}
(s_3, f_1) \models\, & \Box_{\mathrm{F}} \mathrm{t}(No) \wedge \\
& \Box_{\mathrm{F}} \Box_{\mathrm{I}} \big((\neg sa \wedge \neg pe) \rightarrow \mathrm{t}(No) \big) \wedge \\
& \Box_{\mathrm{F}} \Box_{\mathrm{I}} \big(pe \rightarrow \mathrm{t}(Yes) \big)
\end{aligned}
$$

Axiomatics for Atm_0 finite

$\blacksquare \in \{\Box_{\mathtt{I}}, \Box_{\mathtt{F}}\}$

$$
\begin{array}{ll}
\big(\blacksquare\varphi \wedge \blacksquare(\varphi \to \psi)\big) \to \blacksquare\psi & (\mathbf{K}_{\blacksquare}) \\
\blacksquare\varphi \to \varphi & (\mathbf{T}_{\blacksquare}) \\
\blacksquare\varphi \to \blacksquare\blacksquare\varphi & (\mathbf{4}_{\blacksquare}) \\
\neg\blacksquare\varphi \to \blacksquare\neg\blacksquare\varphi & (\mathbf{5}_{\blacksquare}) \\
\Box_{\mathtt{F}}\Box_{\mathtt{I}}\varphi \leftrightarrow \Box_{\mathtt{I}}\Box_{\mathtt{F}}\varphi & (\mathbf{Comm}) \\
\bigvee_{x \in Val} \mathsf{t}(x) & (\mathbf{AtLeast}_{\mathsf{t}(x)}) \\
\mathsf{t}(x) \to \neg\mathsf{t}(y) \text{ if } x \neq y & (\mathbf{AtMost}_{\mathsf{t}(x)}) \\
\big(\mathsf{cn}_{X,Atm_0} \wedge \mathsf{t}(x)\big) \to \Box_{\mathtt{I}}\big(\mathsf{cn}_{X,Atm_0} \to \mathsf{t}(x)\big) & (\mathbf{Funct}) \\
p \to \Box_{\mathtt{F}} p & (\mathbf{Indep}_{\Box_{\mathtt{F}},p}) \\
\neg p \to \Box_{\mathtt{F}}\neg p & (\mathbf{Indep}_{\Box_{\mathtt{F}},\neg p}) \\
\dfrac{\varphi}{\blacksquare\varphi} & (\mathbf{Nec}_{\blacksquare})
\end{array}
$$

$\Rightarrow$ Satisfiability checking: polynomial

Axiomatics for Atm_0 infinite

$\blacksquare \in \{\Box_{\mathtt{I}}, \Box_{\mathtt{F}}\}$

$(\blacksquare\varphi \wedge \blacksquare(\varphi \to \psi)) \to \blacksquare\psi$	$(\mathbf{K}_{\blacksquare})$
$\blacksquare\varphi \to \varphi$	$(\mathbf{T}_{\blacksquare})$
$\blacksquare\varphi \to \blacksquare\blacksquare\varphi$	$(\mathbf{4}_{\blacksquare})$
$\neg\blacksquare\varphi \to \blacksquare\neg\blacksquare\varphi$	$(\mathbf{5}_{\blacksquare})$
$\Box_{\mathtt{F}}\Box_{\mathtt{I}}\varphi \leftrightarrow \Box_{\mathtt{I}}\Box_{\mathtt{F}}\varphi$	$(\mathbf{Comm})$
$\bigvee_{x \in Val} \mathsf{t}(x)$	$(\mathbf{AtLeast}_{\mathsf{t}(x)})$
$\mathsf{t}(x) \to \neg\mathsf{t}(y)$ if $x \neq y$	$(\mathbf{AtMost}_{\mathsf{t}(x)})$
~~$(\mathsf{cn}_{X,Atm_0} \wedge \mathsf{t}(x)) \to \Box_{\mathtt{I}}(\mathsf{cn}_{X,Atm_0} \to \mathsf{t}(x))$~~	~~$(\mathbf{Funct})$~~
$p \to \Box_{\mathtt{F}} p$	$(\mathbf{Indep}_{\Box_{\mathtt{F}},p})$
$\neg p \to \Box_{\mathtt{F}}\neg p$	$(\mathbf{Indep}_{\Box_{\mathtt{F}},\neg p})$
$\dfrac{\varphi}{\blacksquare\varphi}$	$(\mathbf{Nec}_{\blacksquare})$

$\Rightarrow$ Satisfiability checking: in NEXPTIME

Idea of the proof: polynomial reduction into satisfiability checking for product modal logic $S5^2$

From objective to subjective explanation

	Local	Global
Objective	$\texttt{AXp}(\lambda, x)$	$\texttt{PImp}(\lambda, x)$
Subjective	$\texttt{SubAXp}(\lambda, x)$	$\texttt{SubPImp}(\lambda, x)$

Table: Objective vs subjective explanation

$$\texttt{SubPImp}(\lambda, x) =_{def} \Box_{\texttt{F}} \texttt{PImp}(\lambda, x)$$
$$\texttt{SubAXp}(\lambda, x) =_{def} \Box_{\texttt{F}} \texttt{AXp}(\lambda, x)$$

A negative property

Bob

	f_1
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	No
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	Yes
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

	f_2
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	No
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	No
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

PSR principle does not hold in the "black box" setting:

$$(s_3, f_1) \models \Box_{\mathbb{F}}\mathtt{t}(No) \wedge \neg \bigvee_{\lambda \in Term} \mathtt{SubAXp}(\lambda, No)$$

Extension: acquiring information about actual classifier

⇒ **Language**:

$$\varphi \quad ::= \quad p \mid \mathsf{t}(x) \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_{\mathtt{I}}\varphi \mid \Box_{\mathtt{F}}\varphi \mid [\varphi!]\psi$$

$[\varphi!]\psi \approx$ "ψ holds after having discarded all classifiers that do not globally satisfy property φ"

⇒ **Semantic interpretation** of dynamic modality $[\varphi!]$:

$$(\Gamma, s, f) \models [\varphi!]\psi \quad \Longleftrightarrow \quad \text{if } (\Gamma, s, f) \models \Box_{\mathtt{I}}\varphi \text{ then } (\Gamma^{\varphi!}, s, f) \models \psi$$

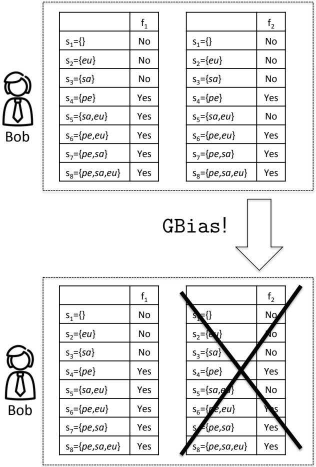
where $\Gamma^{\varphi!} = (S^{\varphi!}, \Phi^{\varphi!})$ is the MCM such that:

$$S^{\varphi!} = S$$

$$\Phi^{\varphi!} = \{f' \in \Phi : \forall s' \in S,\ (\Gamma, s', f') \models \varphi\}$$

Extension: acquiring information about actual classifier

Let PF $= \{eu\}$



	f_1
$s_1=\{\}$	No
$s_2=\{eu\}$	No
$s_3=\{sa\}$	No
$s_4=\{pe\}$	Yes
$s_5=\{sa,eu\}$	Yes
$s_6=\{pe,eu\}$	Yes
$s_7=\{pe,sa\}$	Yes
$s_8=\{pe,sa,eu\}$	Yes

	f_2
$s_1=\{\}$	No
$s_2=\{eu\}$	No
$s_3=\{sa\}$	No
$s_4=\{pe\}$	Yes
$s_5=\{sa,eu\}$	No
$s_6=\{pe,eu\}$	Yes
$s_7=\{pe,sa\}$	Yes
$s_8=\{pe,sa,eu\}$	Yes

GBias!

	f_1
$s_1=\{\}$	No
$s_2=\{eu\}$	No
$s_3=\{sa\}$	No
$s_4=\{pe\}$	Yes
$s_5=\{sa,eu\}$	Yes
$s_6=\{pe,eu\}$	Yes
$s_7=\{pe,sa\}$	Yes
$s_8=\{pe,sa,eu\}$	Yes

	f_2
$s_1=\{\}$	No
$s_2=\{eu\}$	No
$s_3=\{sa\}$	No
$s_4=\{pe\}$	Yes
$s_5=\{sa,eu\}$	No
$s_6=\{pe,eu\}$	Yes
$s_7=\{pe,sa\}$	Yes
$s_8=\{pe,sa,eu\}$	Yes

Bob learns that the classifier is biased thereby being able to conclude that unsuccess of his application is *(abductively) explained* by $\neg pe \wedge \neg eu$

$$(s_3, f_1) \models [\texttt{GBias!}]\texttt{SubAXp}(\neg pe \wedge \neg eu, No)$$

Axiomatics

Axiomatics for the static setting *plus* the following valid equivalences:

$$[\varphi!]p \leftrightarrow (\Box_{\mathtt{I}}\varphi \rightarrow p)$$
$$[\varphi!]\mathsf{t}(x) \leftrightarrow \big(\Box_{\mathtt{I}}\varphi \rightarrow \mathsf{t}(x)\big)$$
$$[\varphi!]\neg\psi \leftrightarrow (\Box_{\mathtt{I}}\varphi \rightarrow \neg[\varphi!]\psi)$$
$$[\varphi!](\psi_1 \wedge \psi_2) \leftrightarrow \big([\varphi!]\psi_1 \wedge [\varphi!]\psi_2\big)$$
$$[\varphi!]\Box_{\mathtt{I}}\psi \leftrightarrow (\Box_{\mathtt{I}}\varphi \rightarrow \Box_{\mathtt{I}}[\varphi!]\psi)$$
$$[\varphi!]\Box_{\mathtt{F}}\psi \leftrightarrow (\Box_{\mathtt{I}}\varphi \rightarrow \Box_{\mathtt{F}}[\varphi!]\psi)$$

and the following rule of replacement of equivalents:

$$\frac{\varphi_1 \leftrightarrow \varphi_2}{\psi \leftrightarrow \psi[\varphi_1/\varphi_2]}$$

$\Rightarrow$ Decidability via the reduction axioms

Outline

1. Explanations in "white box" classifiers
2. Explanations in "black box" classifiers
3. Open problems and future extensions

Open problems

- Exact complexity of satisfiability checking for the logic of "black box" classifiers
- Identify interesting NP fragments:
  - Bounding modal depth
  - Single alternation of $\Box_{\mathtt{F}}/\Diamond_{\mathtt{F}}$ and $\Box_{\mathtt{I}}/\Diamond_{\mathtt{I}}$ modalities sufficient for defining subjective explanation
- Complexity of dynamic extension

Deontic extension

Definition

A multi-classifier model with ideality (MCMI) is a triple $\Gamma = (S, \Phi, \preceq)$ with (S, Φ) an MCM and $\preceq$ a partial preorder on Φ.

$f \preceq f'$: classifier f' is at least as good/ideal as classifier f

"Betterness" modality $[\preceq]$ interpreted wrt pointed MCMI (Γ, s, f):

$$(\Gamma, s, f) \models [\preceq]\varphi \iff \forall f' \in \Phi, \text{ if } f \preceq f' \text{ then } (\Gamma, s, f') \models \varphi$$

Expressive power:

- Obligation modality:

$$\texttt{Oblig}\ \varphi =_{def} \Diamond_{\mathsf{F}}[\preceq]\varphi$$

- Prohibition to have biases:

$$\texttt{Oblig}\ \neg\texttt{GBias}$$

Multi-agent generalization and interactive explanation

Ann | Ann

Bob

	f_2
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	No
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	No
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

	f_4
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	Yes
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	No
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

Bob

	f_1
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	No
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	Yes
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

	f_3
s_1={}	No
s_2={*eu*}	No
s_3={*sa*}	Yes
s_4={*pe*}	Yes
s_5={*sa*,*eu*}	Yes
s_6={*pe*,*eu*}	Yes
s_7={*pe*,*sa*}	Yes
s_8={*pe*,*sa*,*eu*}	Yes

Formal semantics: multi-agent belief bases [Lorini, 2020, AIJ]

References

Liu, X., Lorini, E. (forthcoming). A unified logical framework for explanations in classifier systems. *Journal of Logic and Computation*.

Liu, X., Lorini, E., Rotolo, A., Sartor, G. (forthcoming). Modelling and explaining legal case-based reasoners through classifiers. Proceedings of *JURIX 2022*, IOS Press.

Liu, X., Lorini, E. (2022). A logic of "black box" classifier systems. Proceedings of *WOLLIC 2022*, Springer.

Liu, X., Lorini, E. (2021). A logic for binary classifiers and their explanation. Proceedings of *CLAR 2021*, Springer.