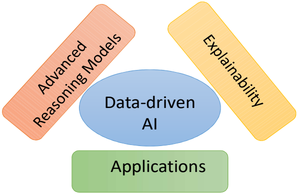
Empowering Data-driven AI by Argumentation and Persuasion: **Main Achievements**

Chair	Leila Amgoud	(CNRS, IRIT)
Co-Chairs	Emiliano Lorini	(CNRS, IRIT)
	Philippe Muller	(UPS, IRIT)

PhD students	Vivien Beuselinck
	Xinghan Liu
	Henri Trenquier

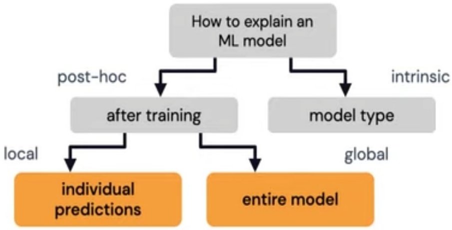
Goals of the chair

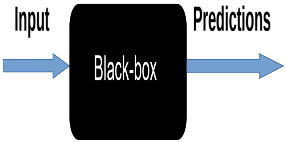


To use advanced logics and argumentation models to explain predictions of machine learning models.

XAI

- ML models carry out predictions
- We want good predictions and know why the model made them
  - Why was the student's application rejected?
  - What can the student do to **change** the situation?
- XAI approaches





Our work

Research questions

1) Which properties should be satisfied by an explanation function?
2) What are the different types of explanations?
3) How to persuade users by those explanations?
4) How to generate explanations in an efficient way?

Contributions

1) Axiomatic foundations of XAI
2) Formal analysis of various types of explanations
3) Dialogical explanations
4) Generation of abductive explanations

1) Axiomatic foundations of XAI

a) Axioms: List of properties that an explainer should satisfy.

- clarify assumptions underlying an explainer
- shed light on weaknesses/strengths of an explainer
- compare different (family of) explainers

Notations

- $\mathbb{E}$: a set of all partial assignments of **values** to **features**
- $\mathcal{X}$: feature space (complete assignments or instances)
- $\mathcal{C}$: a set of classes
- $\kappa : \mathcal{X} \to \mathcal{C}$ a classifier
- **F** : $\mathcal{C} \to \mathcal{P}(\mathbb{E})$ an explainer

1) Axiomatic foundations of XAI (Cont.)

Let $\mathbf{F}$ be an explainer and $x, x' \in \mathcal{C}$

Success	$\mathbf{F}(x) \neq \emptyset$.
Non-Triviality	$\forall\, L \in \mathbf{F}(x),\ L \neq \emptyset$.
Irreducibility	$\forall\, L \in \mathbf{F}(x)$ and $\forall\, t \in L$, $\exists\, I \in \mathbb{X}$, $L \setminus \{t\} \subseteq I$.
Feasibility	$\forall\, L \in \mathbf{F}(x)$, $\exists\, I \in \mathbb{X}_{\mathbf{TR}}(x)$ s.t. $L \subseteq I$.
Representativity	$\forall\, I \in \mathbb{X}_{\mathbf{TR}}(x)$, $\exists L \in \mathbf{F}(x)$ s.t. $L \subseteq I$.
Relevance	$\forall\, L \in \mathbf{F}(x)$ and $\forall\, t \in L$, t is relevant to x.
Coreness	$\forall\, L \in \mathbf{F}(x)$ and $\forall\, t \in L$, t is core to x.
Exhaustivity	$\forall\, t \in \mathtt{Lit}_{\mathbf{T}}$, if t is relevant to x, then $\exists\, L \in \mathbf{F}(x)$, $t \in L$.
Completeness	$\forall\, t \in \mathtt{Lit}_{\mathbf{T}}$, if t is core to x, then $\exists\, L \in \mathbf{F}(x)$, $t \in L$.
Coherence	If $x \neq x'$, then $\forall\, L \in \mathbf{F}(x)$, $\forall\, L' \in \mathbf{F}(x')$, $L \cup L'$ is inconsistent.

1) Axiomatic foundations of XAI (Cont.)

	Vacation	*Concert*	*Meeting*	*Exhibition*	*Hiking*
x_1	0	0	1	0	0
x_2	1	0	0	0	1
x_3	0	0	1	1	0
x_4	1	0	0	1	1
x_5	0	1	1	0	0
x_6	0	1	1	1	0
x_7	1	1	0	1	1

- $\{(V, 0)\} \in \mathbf{F}(0)$
- $\{(M, 0)\} \in \mathbf{F}(1)$

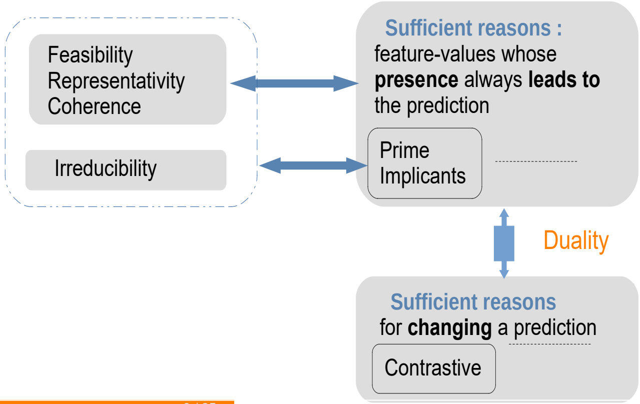
$$\{(V,0),(M,0)\} \text{ is consistent} \Rightarrow \exists I \in \mathbb{X} \text{ s.t. } \{(V,0),(M,0)\} \subseteq I$$

F violates Coherence

1) Axiomatic foundations of XAI (Cont.)

b) Characterization: List of properties that uniquely defines a function

Under the whole feature space



1) Axiomatic foundations of XAI (Cont.)

b) Characterization: List of properties that uniquely defines a function

Under the **whole** feature space

Feasibility
Relevance
Coreness
Completeness
Success
Representativity

↔

Necessary reasons :
feature-values whose **absence** always leads to avoid a prediction

↕ Duality

Necessary reasons
for **changing** a prediction

1) Axiomatic foundations of XAI (Cont.)

c) Impossibility result

- An explanation function which generates **prime implicants** violates Coherence
  - Provides incorrect explanations
  - Limits of LIME, Anchors
  - Limits of (statistical) approaches

- **No explainer** can generate (a subset of) **prime implicants** and guarantees both existence (Success) and correctness (Coherence) of explanations

! Under a **subset** of the space

2) Formal analysis of various types of explanations

a) Modal logics for modelling explanations of classifier systems

(I) Logic of "white box" classifiers: complete knowledge of the classifier

- Basic operators: Instance-quantifying modality $\Box$
- Types of explanation modelled: abductive, contrastive, counterfactual

Prime implicant $\mathtt{PImp}(\lambda, x) =_{def} \Box\Big(\lambda \rightarrow \big(\mathsf{t}(x) \wedge \bigwedge_{p \in Atm(\lambda)} \langle Atm(\lambda) \setminus \{p\} \rangle \neg \mathsf{t}(x)\big)\Big)$

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

(II) Logic of "black box" classifiers: partial knowledge of the classifier

- Extension: classifier-quantifying modality $\blacksquare$ ($\Rightarrow$ product modal logic)
- Crucial distinction: Objective vs subjective explanation

Subjective abductive explanation $\mathtt{SubAXp}(\lambda, x) =_{def} \blacksquare \mathtt{AXp}(\lambda, x)$

(III) Distance-based semantics for counterfactual conditionals: relativized ($\Box\!\!\rightarrow_X$, with $X \subseteq^{fin} Var$) and unrelativized ($\Box\!\!\rightarrow$)

2) Formal analysis of various types of explanations

a) Modal logics for modelling explanations of classifier systems: **Key results**

- Complexity of satisfiability

	"White box"	**"Black box"**
Finite fixed prop. variables	Polynomial	Polynomial
Infinite prop. variables	NP-complete	EXPTIME-hard, in NEXPTIME

- Proof theories

- **"Inexpressibility" result** for counterfactual conditionals:

  In the infinite variable case, the language of unrelativized counterfactual conditionals is not expressive enough to distinguish the concrete distance-based from the abstract similarity-based semantics.

2) Formal analysis of various types of explanations

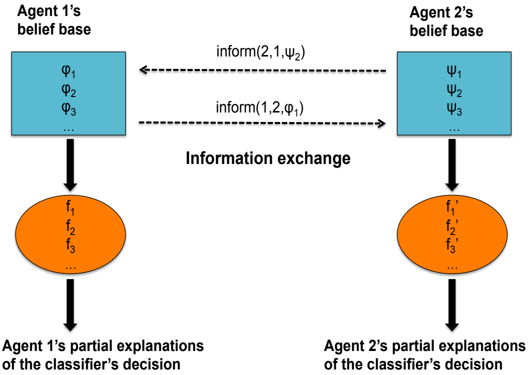
b) Modals logics for causal reasoning

- Opening the box: classifier with internal layers/nodes $\approx$ causal model
- Languages with increasing expressiveness:
  - Causal necessity/possibility
  - Interventionist conditionals
  - Causal counterfactual conditionals
- Modelled notions: actual cause, causal explanation
- Conceptual contribution: novel rule-based semantics for causal reasoning
- Complexity results: satisfiability checking (ST) and model checking (MC)
  - ST and MC for causal necessity is NP-complete
  - MC for interventionist conditionals is in Δ_2^P
- Algorithmic aspects: Reduction of MC to SAT and SAT-based algorithm

3) Dialogical explanations

Rich logical language suitable for representing

- various types of explanations (abductive, contrastive, ...)
- interactive explanations (multi-agent dynamic epistemic setting)



4) Generation of abductive explanations

A weak abductive explanation (wAXp) for $\kappa(x)$ is $E \in \mathbb{E}$ s.t.

i) $E \subseteq x$ ii) $\forall y \in \mathcal{X}$ s.t. $E \subseteq y$, $\kappa(y) = \kappa(x)$

An abductive explanation (AXp) is a subset-minimal wAXp.

There are often constraints on feature space

Integrity constraints $\rightsquigarrow$ **non-feasible** instances

- Years of work < age
- No two distinct students may have the same ID card value

Dependency constraints $\rightsquigarrow$ **dependencies** between assignments

- Social security number $\rightarrow$ surname

4) Generation of abductive explanations (Cont.)

Superfluous explanations

- $\kappa_1(x) = \neg f_1$ $\qquad f_1 \wedge \neg f_2 \rightarrow \bot$
- $x = \langle (f_1, 0), (f_2, 0) \rangle$ $\qquad \kappa_1(x) = 1$
- $E_1 = \{(f_1, 0)\}$ $\qquad E_2 = \{(f_2, 0)\}$

Redundant explanations

- $\kappa_2(x) = f_1 \vee f_2$ $\qquad f_2 \rightarrow f_1$
- $x = \langle (f_1, 1), (f_2, 1) \rangle$ $\qquad \kappa_2(x) = 1$
- $E_1 = \{(f_1, 1)\}$ $\qquad E_2 = \{(f_2, 1)\}$

Exponential number of explanations

- $\kappa_3(x) = f_n$ $\qquad f_n \equiv (\sum_{i=1}^{n-1} f_i \geq \lfloor n/2 \rfloor)$
- $x = \{(f_i, 1) \mid i = 1, \ldots, n\}$ $\qquad \kappa_3(x) = 1$
- $\binom{n}{k}$ *AXp's, where* $k = \lfloor \frac{n}{2} \rfloor$*: all size-k subsets of*

$$\{(f_i, 1) \mid i = 1, \ldots, n-1\}$$

and $\{(f_n, 1)\}$

4) Generation of abductive explanations (Cont.)

a) Three novel types of abductive explanations under constraints

A coverage-based PI-explanation of $\kappa(x)$, with $x \in \mathbb{F}[\mathcal{C}]$, is any $E \in \mathbb{E}$ s.t.

1) $E \subseteq x$,
2) $\forall y \in \mathbb{F}[\mathcal{C}].\ ((E \subseteq y) \rightarrow (\kappa(y) = \kappa(x)))$,
3) $\nexists E' \in \mathbb{E}$ s.t E' satisfies 1) and 2) and strictly subsumes E in $\mathbb{F}[\mathcal{C}]$.

Superfluous explanations

$f_1 \wedge \neg f_2 \rightarrow \perp$

- $\kappa_1(x) = \neg f_1$
- $x = \langle (f_1, 0), (f_2, 0) \rangle$
- $E_1 = \{(f_1, 0)\}$ $\quad \cancel{E_2 = \{(f_2, 0)\}}$ $\quad \cancel{E_3 = \{(f_1, 0), (f_2, 0)\}}$
- $\text{cov}_{\mathbb{F}[\mathcal{C}]}(E_2) = \text{cov}_{\mathbb{F}[\mathcal{C}]}(E_3) \subset \text{cov}_{\mathbb{F}[\mathcal{C}]}(E_1)$

4) Generation of abductive explanations (Cont.)

a) Three novel types of abductive explanations under constraints

Redundant explanations

- $\kappa_2(x) = f_1 \vee f_2$ $\qquad f_2 \rightarrow f_1$
- $x = \langle (f_1, 1), (f_2, 1) \rangle$
- $E_1 = \{(f_1, 1)\}$ $\qquad \cancel{E_2 = \{(f_2, 1)\}}$ $\qquad \cancel{E_3 = \{(f_1, 1), (f_2, 1)\}}$

Exponential number of explanations

- $\kappa_3(x) = f_n$ $\qquad f_n \equiv (\sum_{i=1}^{n-1} f_i \geq \lfloor n/2 \rfloor)$
- $x = \{(f_i, 1) \mid i = 1, \ldots, n\}$
- *The* $\cancel{\binom{n}{k}}$ *AXp's are discarded*
- $\kappa_3(x)$ *has a* ***single CPI-Xp***: $\{(f_n, 1)\}$

4) Generation of abductive explanations (Cont.)

a) Three novel types of abductive explanations under constraints

Minimal coverage-based PI-explanation

A minimal coverage-based PI-explanation of $\kappa(x)$ is a subset-minimal CPI-Xp of $\kappa(x)$.

Preferred coverage-based PI-explanation

A preferred coverage-based PI-explanation of $\kappa(x)$ is a **representative** of the set of miminal CPI-Xp's of $\kappa(x)$.

4) Generation of abductive explanations (Cont.)

a) Three novel types of abductive explanations under constraints

Explanation	Complexity of testing	Complexity of finding one
wAXp	co-NP-complete	polytime
AXp	P^{NP}	$\mathrm{FP}^{\mathrm{NP}}$
CPI-Xp	Π_2^P-complete	$\mathrm{FP}^{\Sigma_2^{\mathrm{P}}}$
mCPI-Xp	Π_2^P-complete	$\mathrm{FP}^{\Sigma_2^{\mathrm{P}}}$
pCPI-Xp	Π_2^P-complete	$\mathrm{FP}^{\Sigma_2^{\mathrm{P}}}$

$\mathrm{FP}^{\mathcal{L}}$ is the class of function problems that can be solved by a polynomial number of calls to an oracle for the language $\mathcal{L}$.

4) Generation of abductive explanations (Cont.)

a) Three novel types of abductive explanations under constraints

$\mathcal{T}$: a sample of **feasible** instances

dCPI-Xp

A dataset-based CPI of $\kappa(x)$ is any $E \in \mathbb{E}$ such that:

1) $E \subseteq x$,
2) $\forall y \in \mathcal{T}.\ ((E \subseteq y) \rightarrow (\kappa(y) = \kappa(x)))$,
3) $\nexists E' \in \mathbb{E}$ *s.t* E' *satisfies 1) and 2) and strictly subsumes* E *in* $\mathcal{T}$.

4) Generation of abductive explanations (Cont.)

Explanation	Complexity of testing	Complexity of finding one
dwAXp	P	polytime
dAXp	P	polytime
dCPI-Xp	P	polytime
dmCPI-Xp	P	polytime
dpCPI-Xp	P	polytime
wAXp	co-NP-complete	polytime
AXp	P^{NP}	$\mathrm{FP}^{\mathrm{NP}}$
CPI-Xp	Π_2^P-complete	$\mathrm{FP}^{\Sigma_2^{\mathrm{P}}}$
mCPI-Xp	Π_2^P-complete	$\mathrm{FP}^{\Sigma_2^{\mathrm{P}}}$
pCPI-Xp	Π_2^P-complete	$\mathrm{FP}^{\Sigma_2^{\mathrm{P}}}$

$\mathrm{FP}^{\mathcal{L}}$ is the class of function problems that can be solved by a polynomial number of calls to an oracle for the language $\mathcal{L}$.

4) Generation of abductive explanations (Cont.)

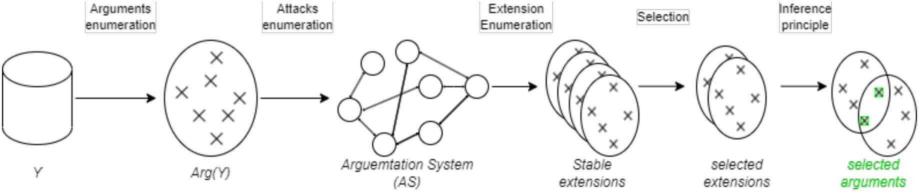
Properties satisfied by each type of explanation

	wAXp	AXp	CPI-Xp	mCPI-Xp	pCPI-Xp	dCPI-Xp	dmCPI-Xp	dpCPI-Xp
Success	✓	✓	✓	✓	✓	✓	✓	✓
Non-Triv.	✓	✓	✓	✓	✓	✓	✓	✓
Irreduc.	×	✓	×	✓	✓	×	✓	✓
Coherence	✓	✓	✓	✓	✓	×	×	×
Consist.	✓	✓	✓	✓	✓	✓	✓	✓
Indep.	×	×	✓	✓	✓	×	×	✓
Non-Equiv.	✓	✓	×	×	✓	×	×	✓

4) Generation of abductive explanations (Cont.)

b) Parameterized family of sample-based explainers

- use argumentation techniques
- guarantee coherence + other axioms
- integrate knowledge
- provide dialogical explanations



Publications (35 A*)

- International Journals (12)
  - 2 AIJ, J. of Approximate Reasoning, J. of Logic and Computation, ...
- International Conferences (53)
  - 11 IJCAI, 7 AAMAS, 3 KR, 3 ECAI, 3 JELIA, 2 AAAI, TARK, ...