FORMAL EXPLAINABILITY

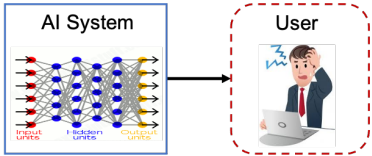
DeepLever Chair

March 25, 2022

Joao Marques-Silva

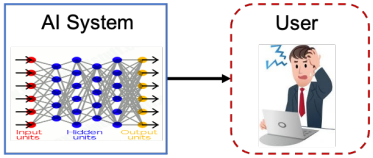
ANITI
Université Fédérale
Toulouse Midi-Pyrénées

XAI & high-risk uses

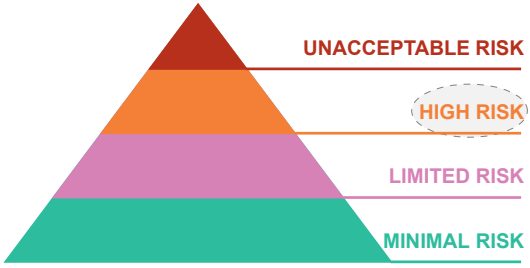


XAI: to help humans understand ML models

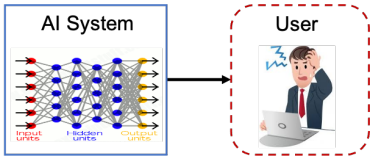
XAI & high-risk uses



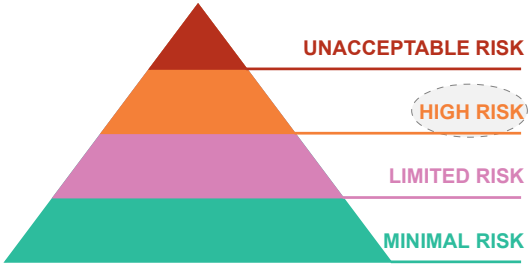
XAI: to help humans understand ML models



XAI & high-risk uses



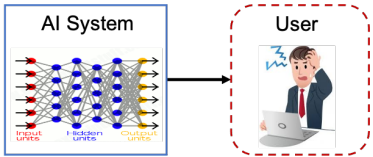
XAI: to help humans understand ML models



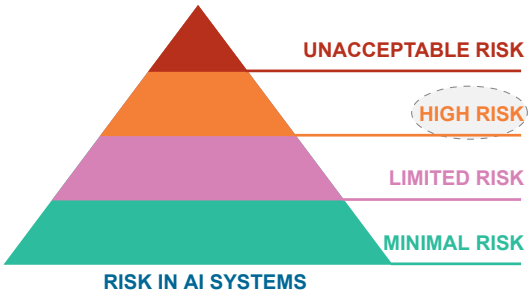
Many examples of high-risk uses: [Pro21]

- Credit worthiness & Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and categorization of people; ...

XAI & high-risk uses



XAI: to help humans understand ML models



Many examples of high-risk uses: [Pro21]

- Credit worthiness & Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and categorization of people; ...

And also safety-critical uses !

Non-formal explanations

- Best-known efforts: model-agnostic and intrinsic interpretability

Non-formal explanations

- Best-known efforts: model-agnostic and intrinsic interpretability
  - **But**: model-agnostic explanations can be **incorrect** & intrinsic interpretability can exhibit **redundancy**

Non-formal explanations

- Best-known efforts: model-agnostic and intrinsic interpretability
  - **But**: model-agnostic explanations can be **incorrect** & intrinsic interpretability can exhibit **redundancy**
  - Examples:

Incorrect explanations (XPs):

Classifier for deciding bank loans

Two samples:

Bessie $:= (v_1, \mathbf{Y})$, Clive $:= (v_2, \mathbf{N})$

Explanation X: age $= 45$, salary $=$ 50K

X is consistent with Bessie $:= (\mathbf{v}_1, \mathbf{Y})$

X is consistent with Clive $:= (\mathbf{v}_2, \mathbf{N})$

$\therefore$ different outcomes **&** same explanation **!?**

Non-formal explanations

- Best-known efforts: model-agnostic and intrinsic interpretability
  - **But**: model-agnostic explanations can be **incorrect** & intrinsic interpretability can exhibit **redundancy**
  - Examples:

Incorrect explanations (XPs):

Classifier for deciding bank loans

Two samples:

Bessie $:= (v_1, \mathbf{Y})$, Clive $:= (v_2, \mathbf{N})$

Explanation X: age $= 45$, salary $= 50$K

X is consistent with Bessie $:= (\mathbf{v}_1, \mathbf{Y})$

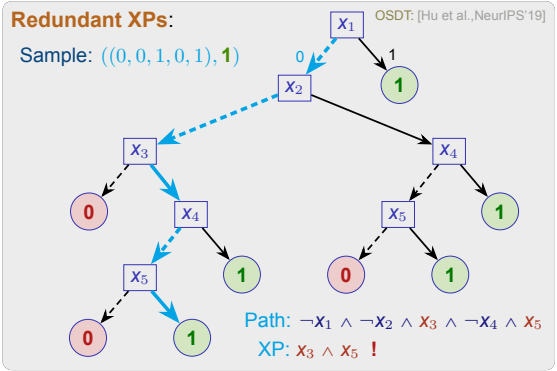
X is consistent with Clive $:= (\mathbf{v}_2, \mathbf{N})$

$\therefore$ different outcomes **&** same explanation **!?**

Redundant XPs:

OSDT: [Hu et al.,NeurIPS'19]

Sample: $((0, 0, 1, 0, 1), \mathbf{1})$



Non-formal explanations

- Best-known efforts: model-agnostic and intrinsic interpretability
  - **But**: model-agnostic explanations can be **incorrect** & intrinsic interpretability can exhibit **redundancy**
  - Examples:

Incorrect explanations (XPs):

Classifier for deciding bank loans

Two samples:

Bessie $:= (v_1, \mathbf{Y})$, Clive $:= (v_2, \mathbf{N})$

Explanation X: age $= 45$, salary $= 50$K

X is consistent with Bessie $:= (\mathbf{v}_1, \mathbf{Y})$

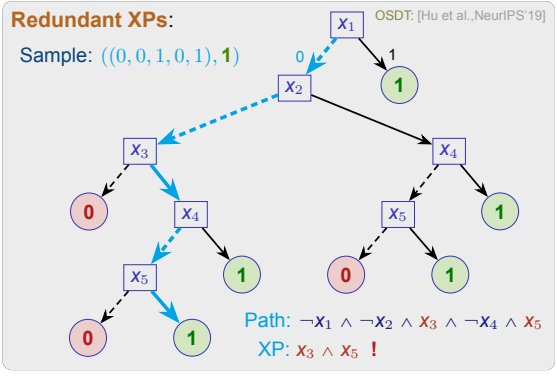
X is consistent with Clive $:= (\mathbf{v}_2, \mathbf{N})$

$\therefore$ different outcomes **&** same explanation **!?**

Redundant XPs:

OSDT: [Hu et al.,NeurIPS'19]

Sample: $((0, 0, 1, 0, 1), \mathbf{1})$



$\therefore$ **more rigorous XAI solutions are vital !**

Formal explanations

Formal XAI in classification:

- Explanations rigorously defined

Formal explanations

Formal XAI in classification:

- Explanations rigorously defined
- Explanation for **Why?** question:
  - Minimal set of features sufficient for ensuring prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{X} \subseteq \mathcal{F}$ s.t.

$$\forall(\mathbf{z} \in \mathbb{F}).\ [\wedge_{i \in \mathcal{X}}(z_i = v_i) \rightarrow (\kappa(\mathbf{z}) = c)]$$

Formal explanations

Formal XAI in classification:

- Explanations rigorously defined
- Explanation for **Why?** question:
  - Minimal set of features sufficient for ensuring prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{X} \subseteq \mathcal{F}$ s.t.

$$\forall(\mathbf{z} \in \mathbb{F}).\ [\wedge_{i \in \mathcal{X}}(z_i = v_i) \rightarrow (\kappa(\mathbf{z}) = c)]$$

- Explanation for **Why Not?** question:
  - Minimal set of features sufficient for changing prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{Y} \subseteq \mathcal{F}$ s.t.

$$\exists(\mathbf{z} \in \mathbb{F}).\ [\wedge_{i \notin \mathcal{Y}}(z_i = v_i) \wedge (\kappa(\mathbf{z}) \neq c)]$$

Formal explanations

Formal XAI in classification:

- Explanations rigorously defined
- Explanation for **Why?** question:
  - Minimal set of features sufficient for ensuring prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{X} \subseteq \mathcal{F}$ s.t.

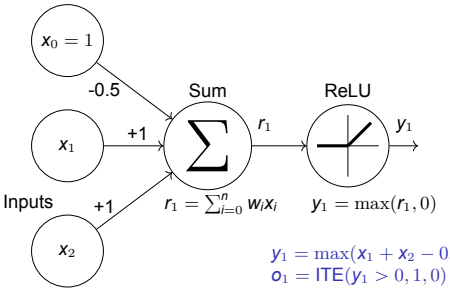
$$\forall(\mathbf{z} \in \mathbb{F}).\ [\wedge_{i \in \mathcal{X}}(z_i = v_i) \rightarrow (\kappa(\mathbf{z}) = c)]$$

- Explanation for **Why Not?** question:
  - Minimal set of features sufficient for changing prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{Y} \subseteq \mathcal{F}$ s.t.

$$\exists(\mathbf{z} \in \mathbb{F}).\ [\wedge_{i \notin \mathcal{Y}}(z_i = v_i) \wedge (\kappa(\mathbf{z}) \neq c)]$$

- Duality results, e.g. between XPs for **Why?** and **Why Not?** questions [INAM20, INM19a]
- More problems: enumeration, membership, preferences, ...

Encoding a simple NN in MILP



x_1	x_2	r_1	y_1	o_1
0	0	-0.5	0	0
1	0	0.5	0.5	1
0	1	0.5	0.5	1
1	1	1.5	1.5	1

MILP encoding:

$x_1 + x_2 - 0.5 = y_1 - s_1$
$z_1 = 1 \rightarrow y_1 \leqslant 0$
$z_1 = 0 \rightarrow s_1 \leqslant 0$
$o_1 = (y_1 > 0)$
$x_1, x_2, z_1, o_1 \in \{0, 1\}$
$y_1, s_1 \geqslant 0$

Instance: $(\mathbf{x}, c) = ((1, 0), 1)$

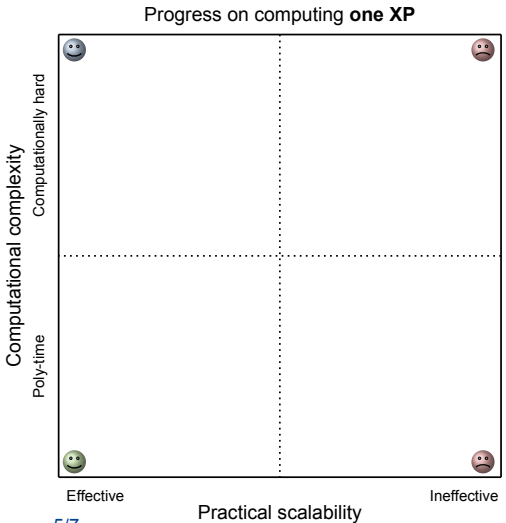
$1 + 0 - 0.5 = 0.5 - 0$
$1 \vee 0.5 \leqslant 0$
$0 \vee 0 \leqslant 0$
$1 = (0.5 > 0)$
$x_1 = 1, x_2 = 0, z_1 = 0, o_1 = 1$
$y_1 = 0.5, s_1 = 0$

Checking: $\mathbf{x} = (0, 0)$

$0 + 0 - 0.5 = 0 - 0.5$
$0 \vee 0 \leqslant 0$
$1 \vee 0.5 \leqslant 0$
$0 = (0 > 0)$
$x_1 = 0, x_2 = 0, z_1 = 1, o_1 = 0$
$y_1 = 0, s_1 = 0.5$

Progress in formal XAI

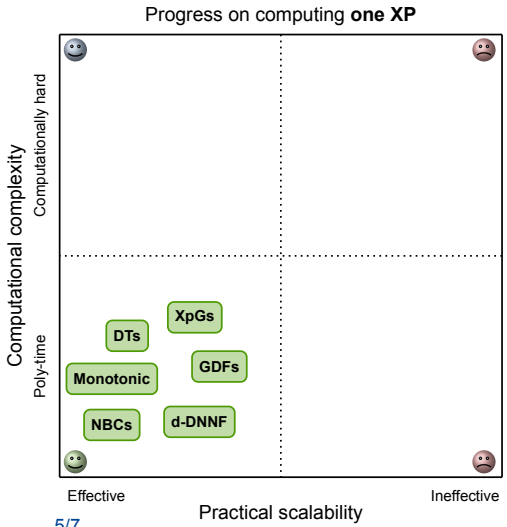
[INM19b, IIM20, MGC+20, MGC+21, HIIM21, IM21, IMS21, CM21, HII+22, IISMS22]



- **Formal XAI efficient for several families of classifiers**

Progress in formal XAI

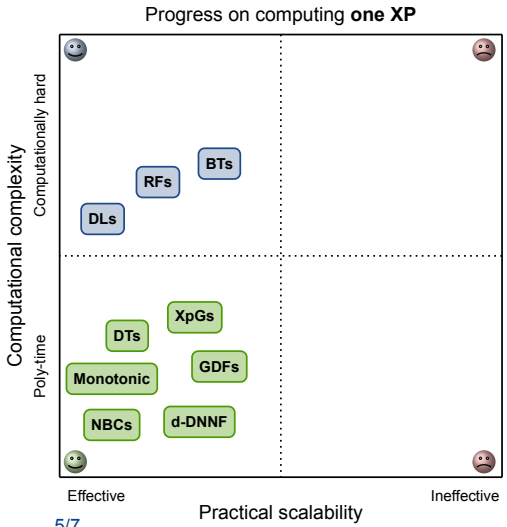
[INM19b, IIM20, MGC+20, MGC+21, HIIM21, IM21, IMS21, CM21, HII+22, IISMS22]



- **Formal XAI efficient for several families of classifiers**
  - Polynomial-time:
    - Naive-Bayes classifiers (NBCs) [MGC+20]
    - Decision trees (DTs) [IIM20, HIIM21]
    - XpG's: DTs, OBDDs, OMDDs, etc. [HIIM21]
    - Monotonic classifiers [MGC+21]
    - Propositional languages (e.g. d-DNNF, ...) [HII+22]
    - Additional results [CM21, HII+22]

Progress in formal XAI

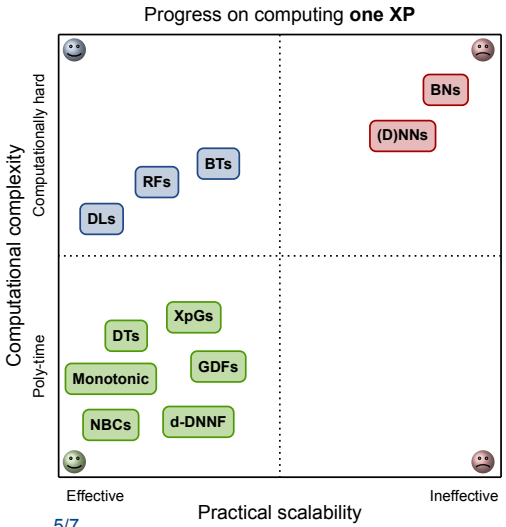
[INM19b, IIM20, MGC+20, MGC+21, HIIM21, IM21, IMS21, CM21, HII+22, IISMS22]



- **Formal XAI efficient for several families of classifiers**
  - Polynomial-time:
    - Naive-Bayes classifiers (NBCs) [MGC+20]
    - Decision trees (DTs) [IIM20, HIIM21]
    - XpG's: DTs, OBDDs, OMDDs, etc. [HIIM21]
    - Monotonic classifiers [MGC+21]
    - Propositional languages (e.g. d-DNNF, ...) [HII+22]
    - Additional results [CM21, HII+22]
  - Comp. hard, but effective (efficient in practice):
    - Random forests (RFs) [IM21]
    - Decision lists (DLs) [IMS21]
    - Boosted trees (BTs) [INM19b, IISMS22]

Progress in formal XAI

[INM19b, IIM20, MGC+20, MGC+21, HIIM21, IM21, IMS21, CM21, HII+22, IISMS22]



- **Formal XAI efficient for several families of classifiers**
  - Polynomial-time:
    - Naive-Bayes classifiers (NBCs) [MGC+20]
    - Decision trees (DTs) [IIM20, HIIM21]
    - XpG's: DTs, OBDDs, OMDDs, etc. [HIIM21]
    - Monotonic classifiers [MGC+21]
    - Propositional languages (e.g. d-DNNF, ...) [HII+22]
    - Additional results [CM21, HII+22]
  - Comp. hard, but effective (efficient in practice):
    - Random forests (RFs) [IM21]
    - Decision lists (DLs) [IMS21]
    - Boosted trees (BTs) [INM19b, IISMS22]
  - Comp. hard, and ineffective (hard in practice):
    - Neural networks (NNs) [INMS19]
    - Bayesian networks (BNs) [SCD18]

Bits and bobs

Scientific output:

XAI/Chair papers	2019(6m)	2020	2021	2022(3m)	Total
CORE A*	1/2	1/6	3/7	4/4	9/19
CORE A	1/1	1/7	2/3	–	4/11
Others	–	1/1	1/2	1/1	3/4
Total	2/3	3/14	6/12	5/5	16/34

More info:

- COALA EU project (w/ industrial partners)
  - XAI demonstrator
- Ongoing patent application on formal XAI
- Joint work with other Chairs:
  - Empowering Data-driven AI by Argumentation and Persuasion (E. Lorini / L. Amgoud)
  - Fair & Robust Learning (N. Asher / J.-M. Loubes)

Research directions

Current lines of research:

- Efficiency of reasoning, e.g. NNs, BNs
- Probabilistic/approximate explanations [APR21, IIN+21, AdLPR22]
  - Reason about probabilistic statement, e.g.

$$\Pr_{\mathbf{z}}(\kappa(\mathbf{z}) = \kappa(\mathbf{v}) \mid \wedge_{i \in \mathcal{X}}(\mathbf{z}_i = \mathbf{v}_i)) \geqslant \delta$$

- Queries related with XPs, e.g. enumeration, membership [HIIM21, HM22]
- Links between XAI and robustness, fairness & model learning [ICS+20]

Q & A

Acknowledgment: joint work with Y. Izza, X. Huang, M. Cooper, N. Asher, A. Ignatiev N. Narodytska, E. Hebrard, M. Siala, et al.

References

Nicholas Asher, Lucas de Lara, Soumya Paul, and Chris Russell.
Counterfactual models for fair and adequate explanations.
In *CD-MAKE*, 2022.

Nicholas Asher, Soumya Paul, and Chris Russell.
Fair and adequate explanations.
In *CD-MAKE*, pages 79–97, 2021.

Martin C. Cooper and Joao Marques-Silva.
On the tractability of explaining decisions of classifiers.
In *CP*, pages 21:1–21:18, 2021.

Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, Martin Cooper, Nicholas Asher, and Joao Marques-Silva.
Tractable explanations for d-DNNF classifiers.
In *AAAI*, February 2022.

References II

Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva.
On efficiently explaining graph-based classifiers.
In *KR*, pages 356–367, 2021.

Xuanxiang Huang and Joao Marques-Silva.
On deciding feature membership in explanations of SDD & related classifiers.
CoRR, abs/2202.07553, 2022.

Alexey Ignatiev, Martin C. Cooper, Mohamed Siala, Emmanuel Hebrard, and Joao Marques-Silva.
Towards formal fairness in machine learning.
In *CP*, pages 846–867, 2020.

Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva.
On explaining decision trees.
CoRR, abs/2010.11034, 2020.

References III

Yacine Izza, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva.
Efficient explanations with relevant sets.
CoRR, abs/2106.00546, 2021.

Alexey Ignatiev, Yacine Izza, Peter J. Stuckey, and Joao Marques-Silva.
Using MaxSAT for efficient explanations of tree ensembles.
In *AAAI*, February 2022.

Yacine Izza and Joao Marques-Silva.
On explaining random forests with SAT.
In *IJCAI*, pages 2584–2591, 2021.

Alexey Ignatiev and Joao Marques-Silva.
SAT-based rigorous explanations for decision lists.
In *SAT*, pages 251–269, 2021.

References IV

Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva.
From contrastive to abductive explanations and back again.
In *AI*IA*, pages 335–355, 2020.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva.
On relating explanations and adversarial examples.
In *NeurIPS*, pages 15857–15867, 2019.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva.
On validating, repairing and refining heuristic ML explanations.
CoRR, abs/1907.02509, 2019.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva.
Abduction-based explanations for machine learning models.
In *AAAI*, pages 1511–1519, 2019.

References V


Joao Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska.
Explaining naive bayes and other linear classifiers with polynomial time and delay.
In *NeurIPS*, 2020.


Joao Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska.
Explanations for monotonic classifiers.
In *ICML*, pages 7469–7479, 2021.


EU Proposal.
European Artificial Intelligence Act – Proposal.
https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206, 2021.

References VI

Andy Shih, Arthur Choi, and Adnan Darwiche.
A symbolic approach to explaining Bayesian network classifiers.
In *IJCAI*, pages 5103–5111, 2018.