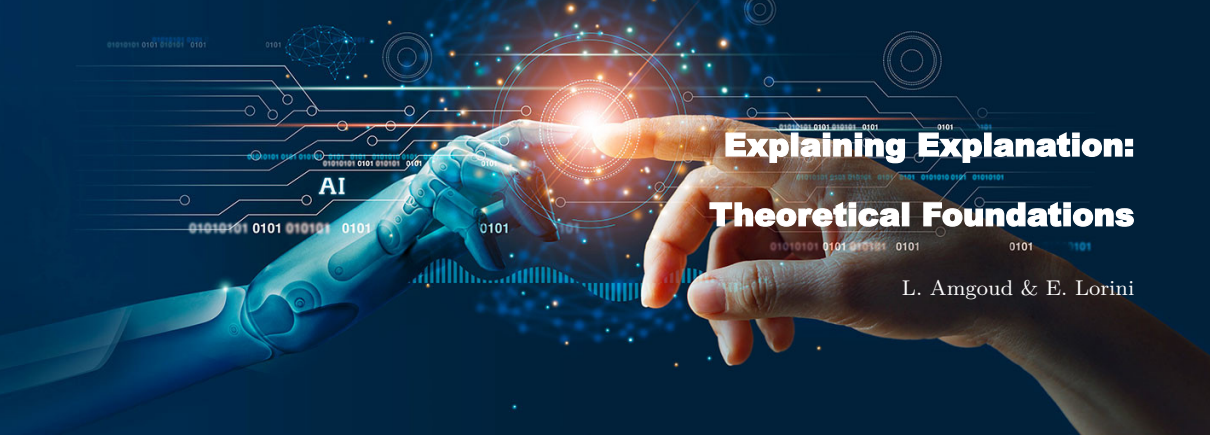


Explaining Explanation: Theoretical Foundations

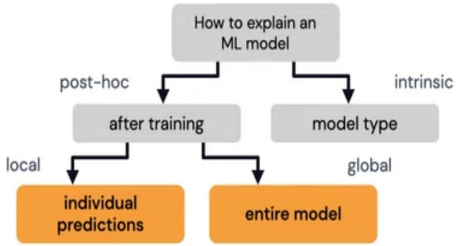
L. Amgoud & E. Lorini

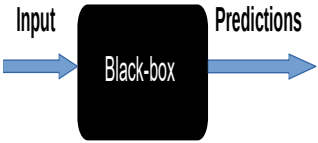
ANITI
Université Fédérale
Toulouse Midi-Pyrénées

What is explainable AI?

- ML models carry out predictions
- We want to have good predictions and also know why the model made them
  - Why was the student's application rejected?
  - What can the student do to change the situation?
- Explanations are important for
  - Fairness: ensure that decisions are based on fair principles
  - Privacy: protect sensitive data
  - Trust: generate trust in the models

How to explain a ML model?





Our work

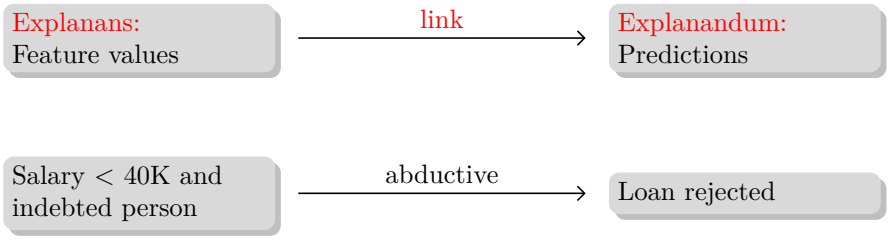
Research questions

1) What makes a good explanation?
2) What are the types of explanations?
3) How to persuade users by those explanations?

Contributions

1) Formal properties of explanation functions
2) Identification of (families of) explanation functions satisfying the properties
3) Logical setting for representing and exchanging various types of explanations

Explanation

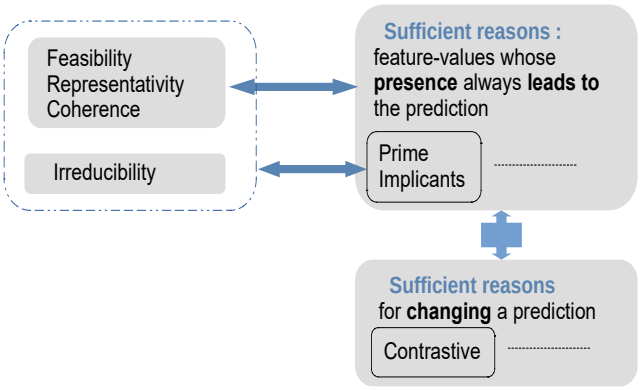


Theoretical foundations of XAI

1) Axioms: (formal properties) that an explanation function should satisfy
   - (Success) Existence of explanations
   - (Coherence) Consistency of a set of explanations
     - I don't hike because I'm not on vacation
     - I hike because I don't have a meeting
     - What if I'm not on vacation and I don't have a meeting?
     - At least one of the two explanations is incorrect
   - + 8 other axioms

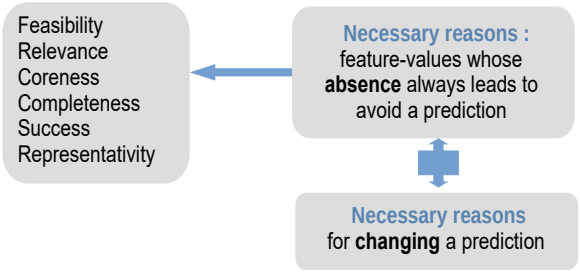
Theoretical foundations of XAI

2) Characterizations: List of properties that uniquely define a function



Theoretical foundations of XAI

2) Characterizations: List of properties that uniquely define a function



3) Impossibility results

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

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

Theoretical foundations of XAI

4) Novel rigorous explanation functions that

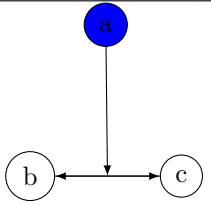
- guarantee correct explanations
- approximate "real" explanations
- satisfy desirable properties
- integrate knowledge
- provide dialogical explanations

$a = \langle \{\neg \text{Vacation} \rightarrow \text{Meeting}\} \rangle$

$b = \langle \{\neg \text{Vacation}\}, \neg \text{Hike} \rangle$

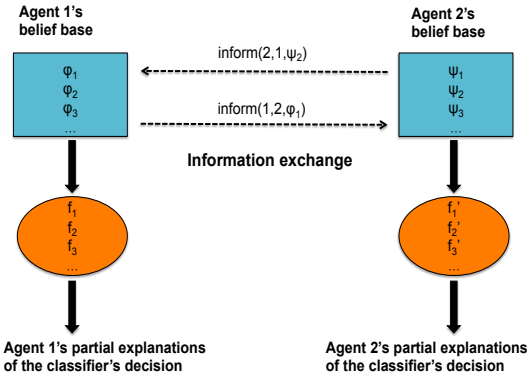
$c = \langle \{\neg \text{Meeting}\}, \text{Hike} \rangle$

Vacation	Concert	Meeting	Exhibition	Hiking
0	0	1	0	0
1	0	0	0	1
0	0	1	1	0
1	0	0	1	1
0	1	1	0	0
0	1	1	1	0
1	1	0	1	1



Modal logic for modelling explanations

- Rich logical language suitable for representing
  - various types of explanations (abductive, constrastive, ...)
  - interactive explanations (multi-agent dynamic epistemic setting)



Modal logic for modelling explanations

- Rich logical language suitable for representing
  - various types of explanations (abductive, constrastive, ...)
  - interactive explanations (multi-agent dynamic epistemic setting)

- Key results
  - Proof theories
  - Complexity of satisfiability
  - Model checking

Summary

- Axioms that explanation functions would satisfy
- Formal characterizations of families of explanation functions satisfying subsets of axioms
- Shedding light on weaknesses/strengths of existing explanation functions
- Novel explainers with good formal properties
- Logical theory of explanations

Some references

- L. Amgoud, D. Doder, S. Vesic. Evaluation of argument strength in weighted graphs: Foundations and semantics. In Artificial Intelligence J., 2022.
- L. Amgoud. Principle-based approach for explainability. Int. J. of Approximate Reasoning, 2022.
- E. Lorini, P. Song. A Computationally Grounded Logic of Awareness. J. of Logic and Computation, 2022.
- L. Amgoud, V. David. A general setting for gradual semantics dealing with similarity. AAAI-2021.
- L. Amgoud, V. Beuselinck. Equivalence of Semantics in Argumentation. In KR'2021.
- L. Amgoud. Explaining Black-box Classification Models with Arguments. ICTAI-2021.
- L. Amgoud, V. Beuselinck. Equivalence of semantics in argumentation. KR-2021.
- J. Luis Fernandez, D. Longin, E. Lorini, F. Maris. A Simple Framework for Cognitive Planning. AAAI-21.
- X. Liu, E. Lorini. A Logic for Binary Classifiers and Their Explanation. CLAR 2021.
- E. Lorini, F. Schwarzentruber. Multi-agent belief base revision. IJCAI-2021.
- E. Lorini, F. Schwarzentruber. A Computationally Grounded Logic of Graded Belief. JELIA-2021.
- L. Amgoud. Evaluation of analogical arguments by Choquet integral. ECAI-2020.
- E. Lorini. Rethinking epistemic logic with belief bases. Artificial Intelligence Journal, 2020.
- L. Amgoud. A Replication Study of Semantics in Argumentation. IJCAI-2019.
- L. Amgoud, D. Doder. Gradual Semantics Accounting for Varied-Strength Attacks. AAMAS-2019.