A general introduction to eXplainable Artificial Intelligence

Equitable Algorithms (EQUAL) final workshop
Bologna, January 23rd, 2026

Marco Lippi
marco.lippi@unifi.it

UNIVERSITÀ
DEGLI STUDI
FIRENZE

UNIVERSITÀ DEGLI STUDI FIRENZE

What is XAI?

[Image from Wikipedia]

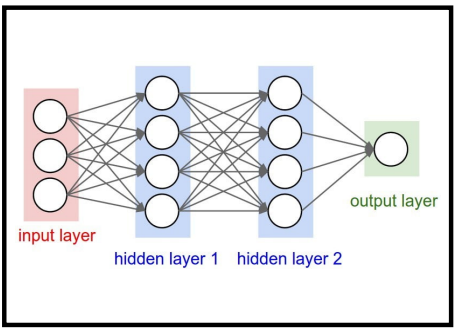
What is XAI?

Explainable AI (XAI) explores and investigates **methods** to produce or complement AI models to make **accessible** and **interpretable** the internal logic and the outcome of algorithms, making such process **understandable by humans**

UNIVERSITÀ DEGLI STUDI FIRENZE

Black-box models

Black-box models



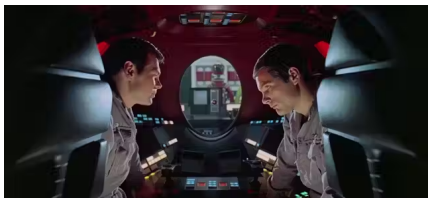
CAT

UNIVERSITÀ DEGLI STUDI FIRENZE

The alignment problem

"If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively… We had better be quite sure that the purpose put into the machine is the purpose which we really desire."

Norbert Wiener, 1960



[Source: 2001: A space odyssey (1968)]

Risks and perils in the (mis)use of AI

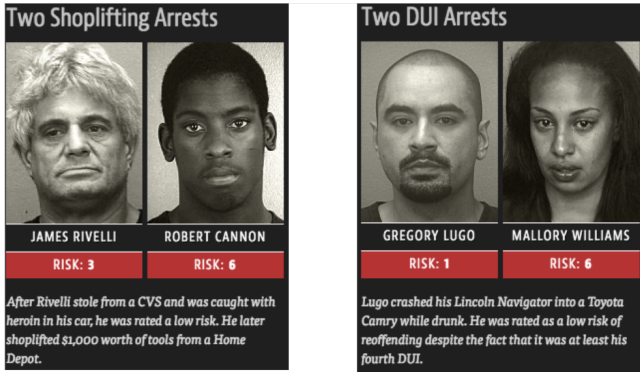
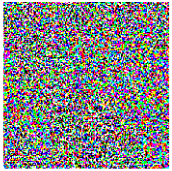


Image source: Propublica

Black-box models can be fooled

$\boldsymbol{x}$
$y =$ “panda”
w/ 57.7% confidence

$+ .007 \times$



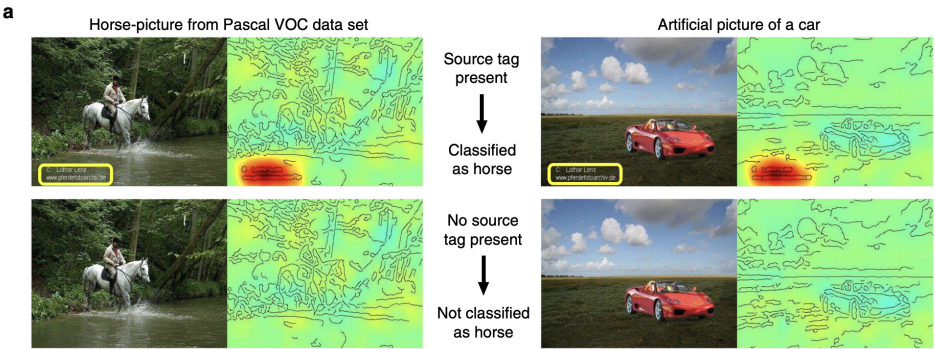
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“nematode”
w/ 8.2% confidence

$=$

$\boldsymbol{x} + \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“gibbon”
w/ 99.3 % confidence

Image source: Goodfellow et al., 2016

Bias in data (e.g., Clever Hans effect)



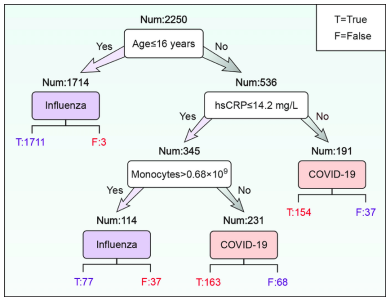
[Source: Lapuschkin et al., 2023]

Bias in data (e.g., leading to discrimination)

Interpretability vs. Explainability

Some models do not need to be **explained**: they are **inherently interpretable** by design

- Decision trees
- Generalized linear models
- Rule-based systems (lists, sets)
- Counterfactual reasoning
- Inductive logic programming
- …



[Image from Zhou et al., 2021]

Interpretability vs. Explainability

Interpretable-by-design approaches

- Sometimes interpretable models are **also** accurate!
- In that case, they should be preferred!
- Use a black-box model only if there is **need** to do it…
- …or if it is the only available model (e.g., a proprietary system)
- When using a black-box model, need to find **post-hoc explanations**

UNIVERSITÀ DEGLI STUDI FIRENZE

Classic approaches

- Linear models
- Decision trees
- Rule-based approaches
- Logic-based approaches

...But there is more than just that!

Optimal decision trees

Existing approaches to decision tree learning are **greedy** as they consider at each split which is the most appropriate attribute (e.g., via Gini index or Information Gain Ratio)

Unfortunately, these approaches fail in optimizing the **overall cost function** of the whole decision tree, and are not designed to optimize any particular performance metric

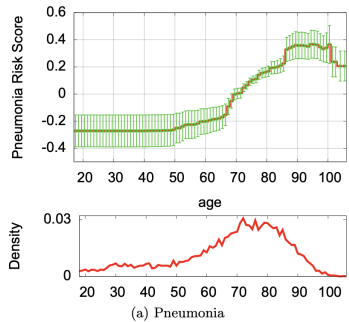
Full decision tree optimization is known to be an **NP-complete** problem

$$\min_{f \in \text{Trees}} \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(f, z_i) + C \cdot \text{NumLeaves}(f)$$

$$s.t. \qquad depth(f) \leq D$$

Generalized additive models

Learn a function **for each feature** and interpret each function as a 1-D chart



[Source: Caruana et al., 2015]

Prototypes: this looks like that...

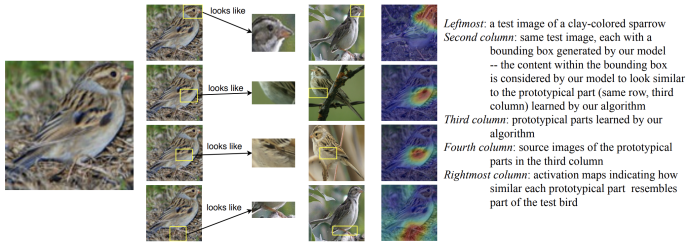


Figure 1: Image of a clay colored sparrow and how parts of it look like some learned prototypical parts of a clay colored sparrow used to classify the bird's species.

[Source: Chen et al., 2019]

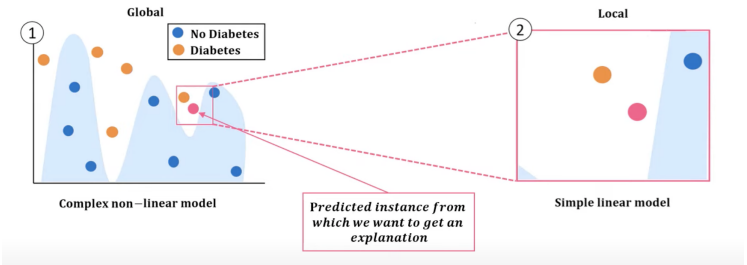
Interpretability vs. Explainability

Black-box explanations

- **Global explanation**: full explanation of an opaque AI system through an interpretable and transparent model that fully captures its logic
- **Local explanation**: does not aim to reconstruct the whole opaque AI system, but to build an explanator that provides explanations for any specific instance

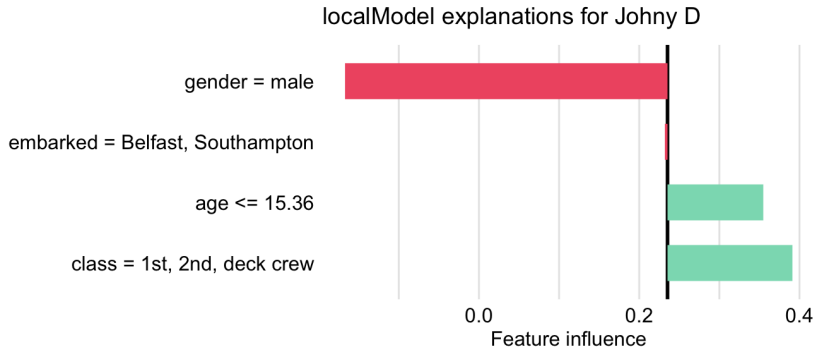
Local Interpretable Model-agnostic Explanations (LIME)

- Look for **local** explanations simpler than global ones
- Learn a **simple** classifier from a set of perturbed examples



[Fonte: Aleix Nieto]

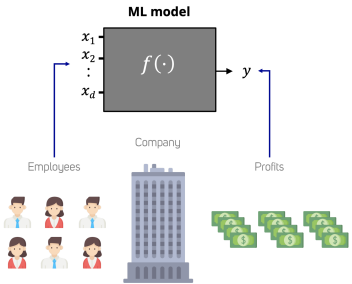
Local Interpretable Model-agnostic Explanations (LIME)



[Fonte: Biecek and Burzykowski, 2022]

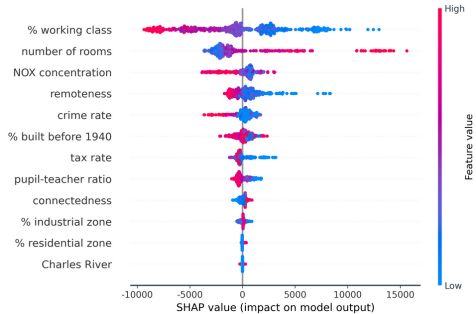
SHapley Additive exPlanations (SHAP)

- Idea from cooperative game theory



[Fonte: Covert and Lee]

SHapley Additive exPlanations (SHAP)



[Fonte: Kaggle]

Attention

Attention is **not explanation** [PDF] arxiv.org

S Jain, BC Wallace - arXiv preprint arXiv:1902.10186, 2019 - arxiv.org

… different **attention** distributions that nonetheless yield equivalent predictions. Our findings show that standard **attention** modules do **not** provide meaningful **explanations** and should **not** …

☆ Salva 99 Cita Citato da 2109 Articoli correlati Tutte e 5 le versioni »

Attention is **not not explanation** [PDF] arxiv.org

S Wiegreffe, Y Pinter - arXiv preprint arXiv:1908.04626, 2019 - arxiv.org

… Existence does **not** Entail Exclusivity. On a more theoretical level, we hold that **attention** scores are used as providing an **explanation**; **not** the **explanation**. The final layer of an LSTM …

☆ Salva 99 Cita Citato da 1496 Articoli correlati Tutte e 9 le versioni »

[Source: Bao et al., 2018]

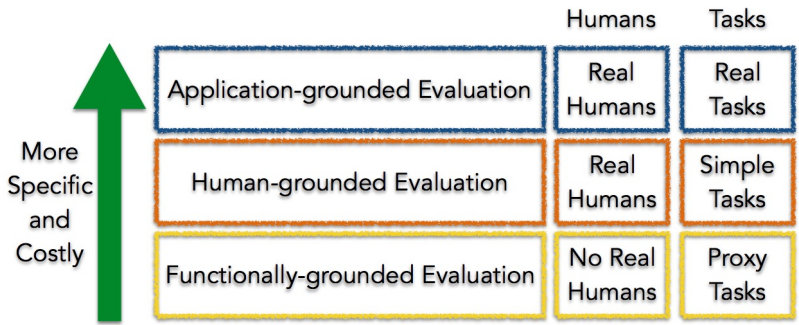
UNIVERSITÀ DEGLI STUDI FIRENZE

Evaluating explainability

Evaluating explainability and interpretability is not straightforward!
There are several ways in which the problem can be addressed…

- Is the system working as designed?
- Are system users treated fairly?
- Is the system compliant to the law?
- Shall we evaluate explanations in the context of the application?
- Shall we evaluate explanations via a quantifiable proxy?
- Are some explanations better than others?

Evaluating explainability



[Figure by Doshi-Velez and Kim]

UNIVERSITÀ DEGLI STUDI FIRENZE

Evaluating explainability

Quantitative performance metrics (with human contribution)

- **Completeness**: whether an explanation is complete for a user
- **Simplicity**: simpler explanations should be preferred (Occam's Razor)
- **Complexity**: needed time for a human being to understand the explanation
- **Plausibility**: whether an explanation is persuasive for a user
- **Simulability**: whether an explanation can be used on new data
- **Relevance**: specific metric for specific domain, such as clinical or juridical

Evaluating explainability

Auxiliary or proxy performance metrics

- **Sensitivity**: to what extent a model is sensible to the value of a feature
- **Continuity**: similar examples in input space should have similar explanations
- **Consistency**: to what extant en explanation is consistent across different models
- **Computational cost**: time required to compute the explanation
- **Correctness**: comparison with some ground-truth

UNIVERSITÀ DEGLI STUDI FIRENZE

The Future of XAI

- Understanding the behavior of AI systems will be more and more important
- Still a crucial element for classifiers working on tabular data
- Beyond the interpretability vs. accuracy trade-off
- How about XAI with Large Language Models?