Explainable AI e sistemi decisionali ibridi

Roberto Pellungrini
Scuola Normale Superiore
Pisa

SCUOLA
NORMALE
SUPERIORE

SOBIGDATA.it
ITALIAN RESEARCH INFRASTRUCTURE

SoBigData

Explainable AI

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

- **Explicability**, understood as incorporating both **intelligibility** *("how does it work?")* for non-experts, e.g., patients or business customers, and for experts, e.g., product designers or engineers) and **accountability** (*"who is responsible for"*).

Interpretability

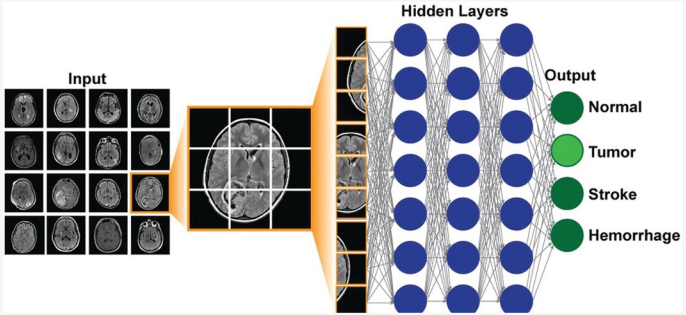
- To *interpret* means to give or provide the meaning or to explain and present in understandable terms some concepts.

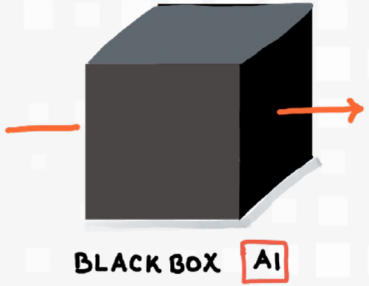
- In data mining and machine learning, interpretability is the *ability to explain* or to provide the meaning *in understandable terms to a human*.

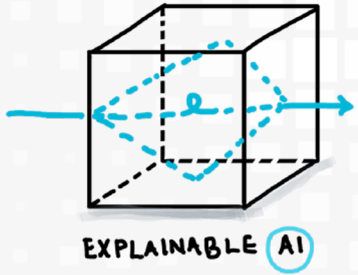
- https://www.merriam-webster.com/

- Finale Doshi-Velez and Been Kim. 2017. ***Towards a rigorous science of interpretable machine learning***. arXiv:1702.08608v2.

Explainable AI







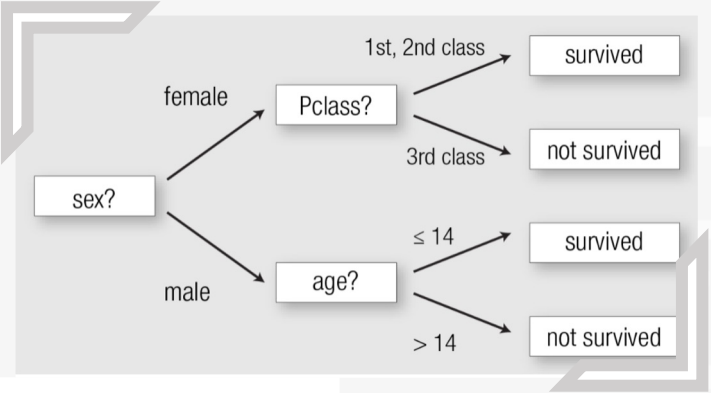
Understand the internal reasoning of the model

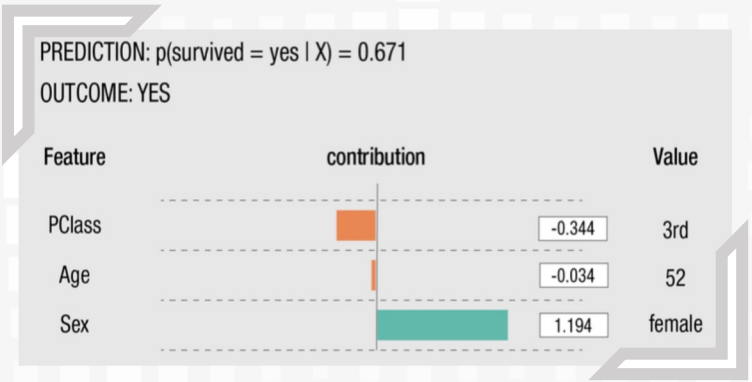
Identify bias, errors and problems related to the model

Develop better models

Increase user's awareness and trust

Recognized Interpretable Models & Explanations

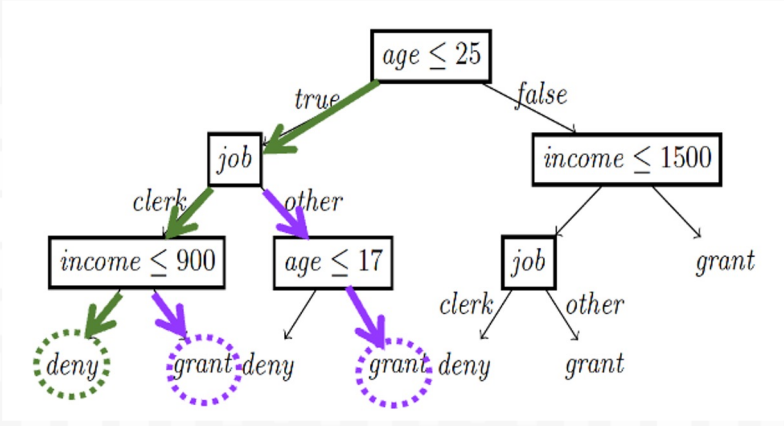




very dark beer . pours a nice finger and a half of creamy foam and stays throughout the beer . major coffee-like taste with hints of chocolate . if you like black coffee , you will love this

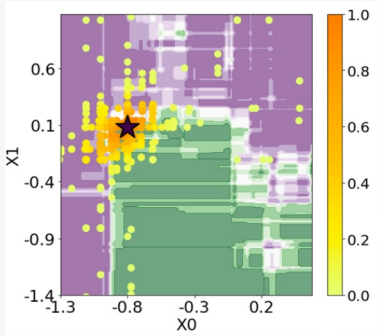
LORE: LOcal Rule-based Explainer

- LORE extends LIME adopting as local surrogate a decision tree classifier and by generating synthetic instances through a genetic procedure that accounts for both instances with the same labels and different ones.
- It can be generalized to work on images and text using the same data representation of LIME.



parent 1	25	clerk	10k	yes
parent 2	30	other	5k	no
↓				
children 1	25	other	5k	yes
children 2	30	clerk	10k	no

parent	25	clerk	10k	yes
↓			↓	
children	27	clerk	7k	yes



r = {age ≤ 25, job = clerk, income ≤ 900} -> deny

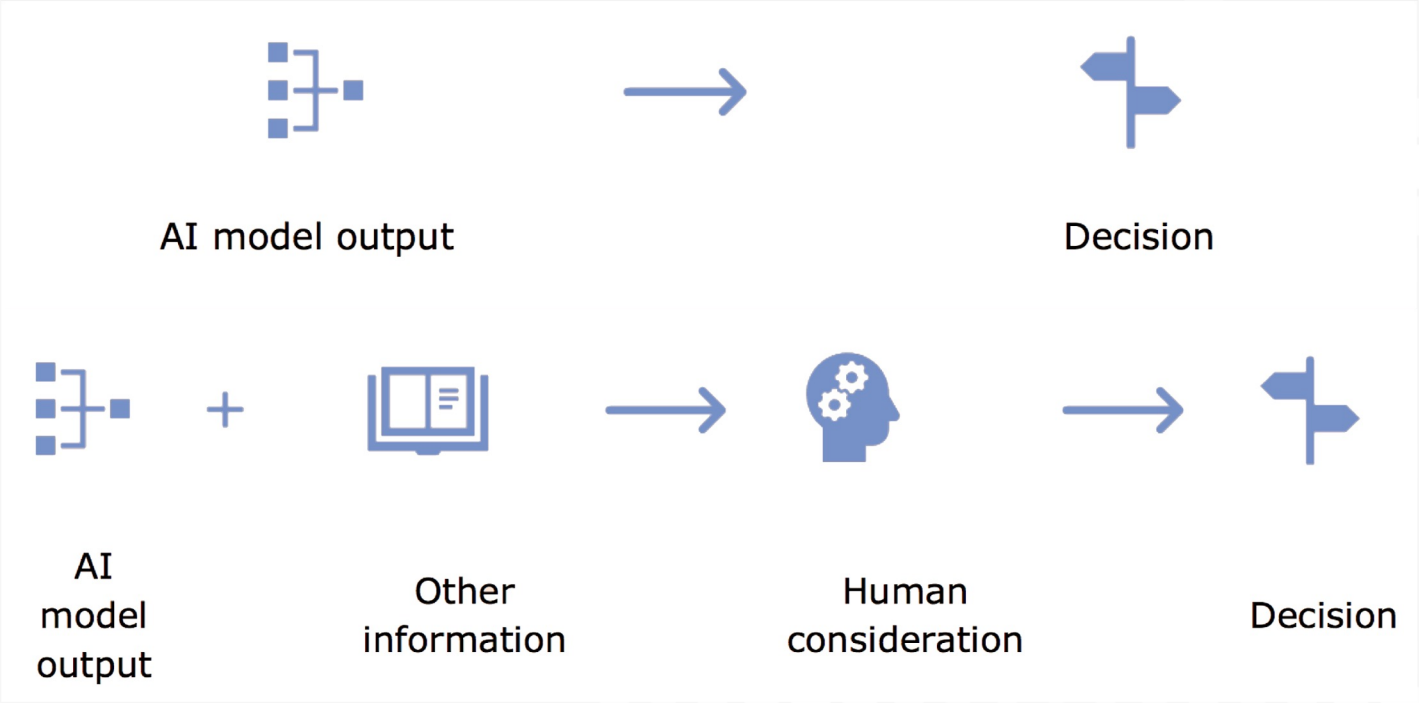
Φ = {({income > 900} -> grant),
({17 ≤ age < 25, job = other} -> grant)}

Guidotti, R., Monreale, A., Ruggieri, S. *et al.* **Stable and actionable explanations of black-box models through factual and counterfactual rules**. *Data Min Knowl Disc* (2022). https://doi.org/10.1007/s10618-022-00878-5

Right of Explanation

Since 25 May 2018, GDPR establishes a right for all individuals to obtain "**meaningful explanations** of the logic involved" when "automated (algorithmic) individual **decision-making**", including profiling, takes place.

What is AI-assisted decision making?



Hybrid Decision Making Systems

Two kinds of **agents**:
humans and machines

A **task** to solve (e.g, problem solving or decision making)

The **joint behavior** of the agents while addressing the task

- Reduce the workload of human experts
- Speed up mechanical processes
- Reduce human bias

Three Learning Paradigms

1

Human Oversight

2

Learning to Abstain

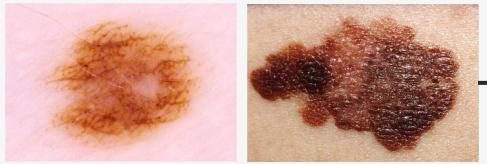
3

Learning Together

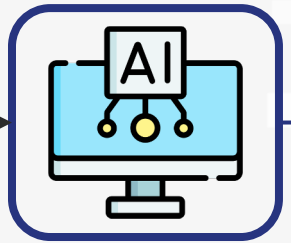
- Punzi C., Pellungrini R., Setzu M., Giannotti F., Pedreschi D.,. **AI, Meet Human: Learning Paradigms for Hybrid Decision Making Systems**, submitted to ACM Computing Surveys, 2023.

Human Oversight

1



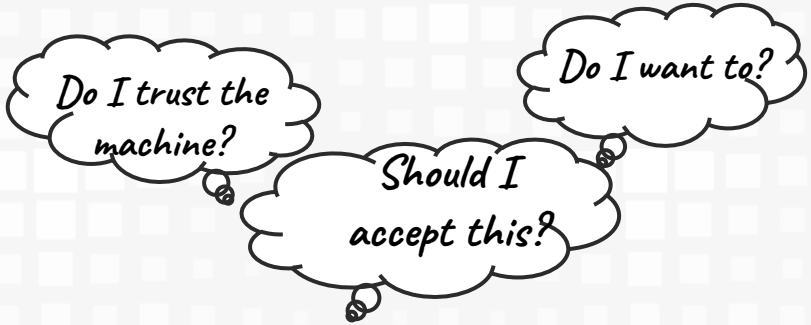
Are these skin lesions melanomas?



Yes

No

Human step



Yes

No

Yes

No

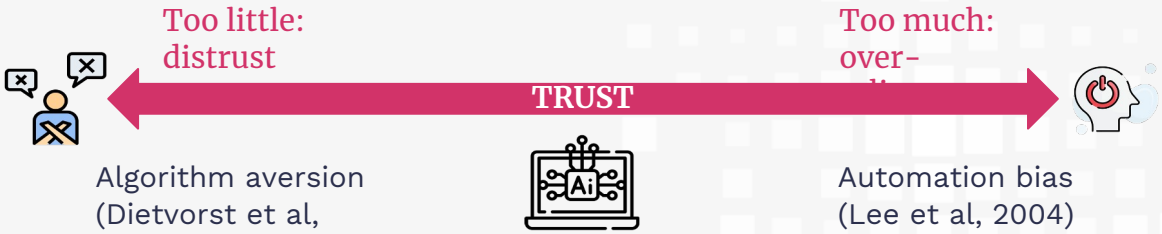
Human Oversight: pros and cons

Pros

- Complete human autonomy

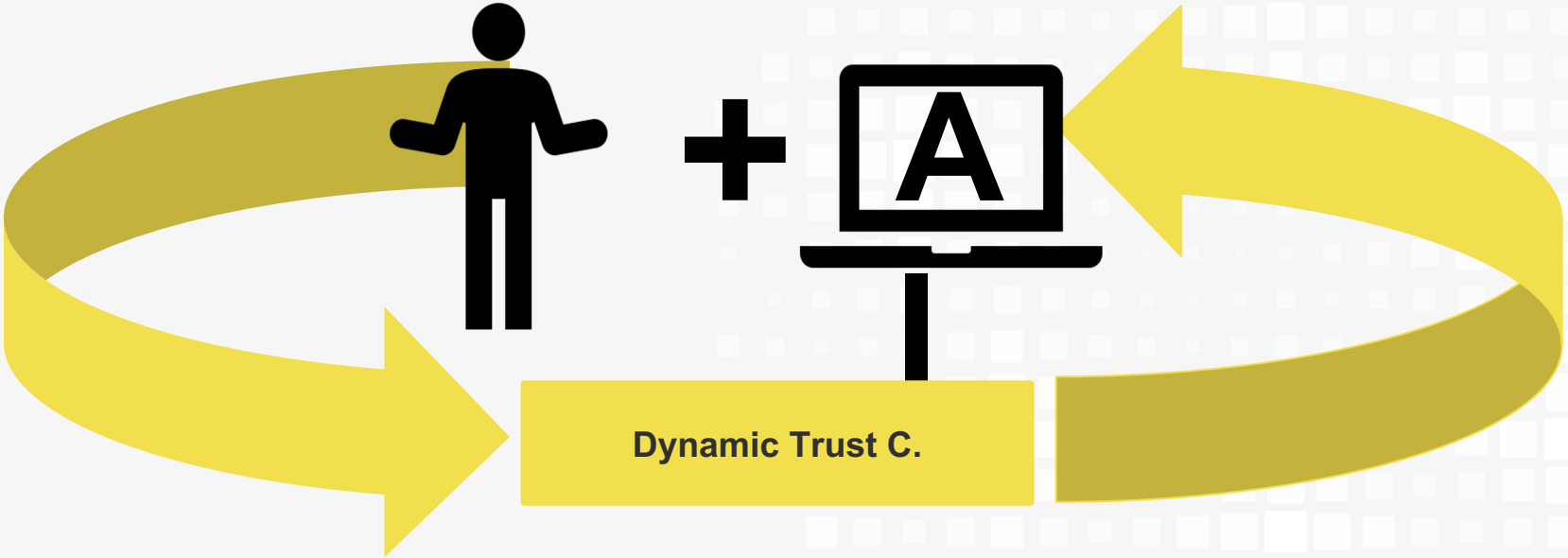
Cons

- Huge, non-reducible human oversight effort
- **Obscure machine reasoning**



Algorithm aversion: people erroneously avoid algorithms after seeing them err, Dietvorst et al.
Human-AI interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice., Alon-Barkat & Busuioc
Dimensions of Diversity in Human Perceptions of Algorithmic Fairness., Grgic-Hlaca et al.
Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making, English et al.
Dice in the Black Box: User Experiences with an Inscrutable Algorithm, Springer et al.

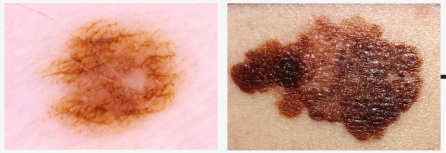
Current Focus of Explainability



Explainability is the first step in developing models able to communicate with a human counterpart, so that decisions are explained and a dynamic trust can be established between the human and the AI.

Learning to Abstain

2



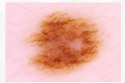
Are these skin lesions melanomas?

abstention policy

I expect the machine to give a better prediction

Machine step

The classifier predicts



No

)

I expect the human to give a better prediction

Human step

Defer to human that predicts autonomously

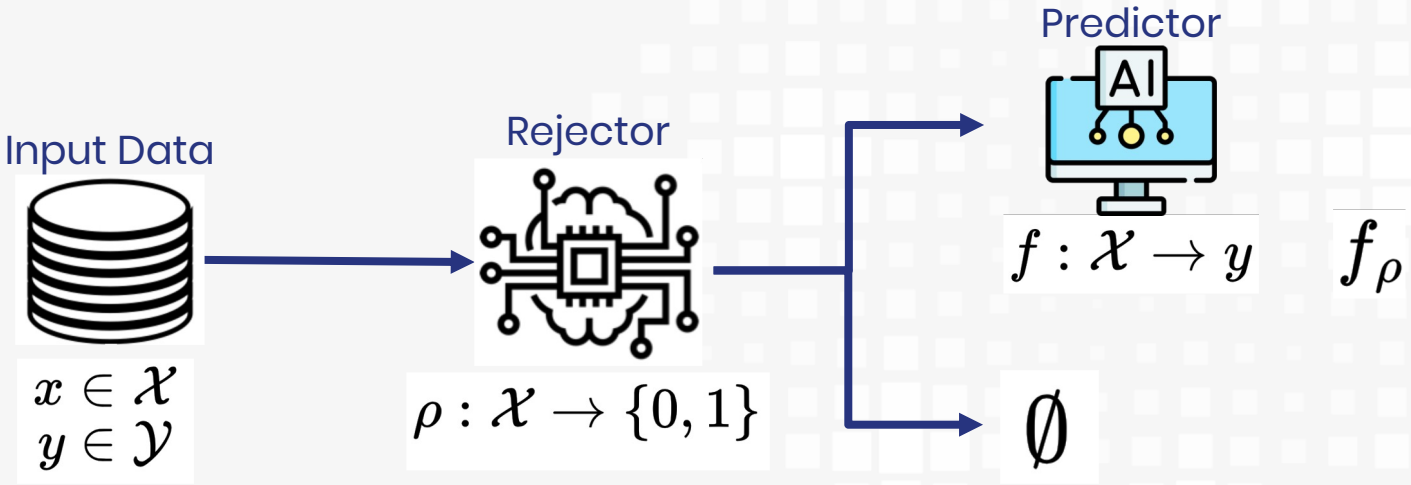
( , )

Yes

SoBigData

Learning to reject (selective classification)

2.1



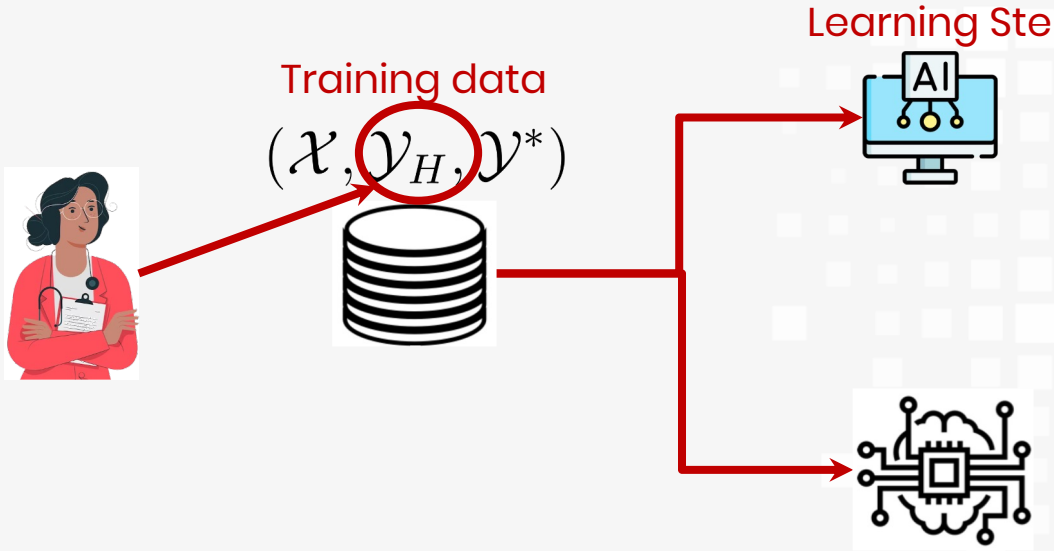
$$f(n) = \begin{cases} C_c, & f_\rho(x) = y \\ C_r, & f_\rho(x) = \emptyset \\ C_e, & f_\rho(x) \notin \{y, \emptyset\} \end{cases} \qquad \mathscr{L}(f_\rho, x, y) := \mathbb{1}_{[\rho(x)=0]} \mathbb{1}_{[f(x) \neq y]} + c \mathbb{1}_{[\rho(x)=1]}$$

C. Chow. 1970. **On optimum recognition error and reject tradeoff.** IEEE Transactions on Information Theory 16, 1(1970), 41–46.

Ni, C., Charoenphakdee, N., Honda, J., and Sugiyama, M. (2019). **On the Calibration of Multiclass Classification with Rejection.** arXiv preprint arXiv:1901.10655, pages 1–31.

Learning to defer

2.
2



The Classifier is learned

$$f : \mathcal{X} \to \mathcal{Y}_M$$

Deferrer is learned

$$\rho : \mathcal{X} \to \{0, 1\}$$

$$\rho(x) = \begin{cases} 0, & \text{if the classifier is "better" than the human} \\ 1, & \text{otherwise} \end{cases}$$

Learning to defer

2.
2

Input data

x

x_1 

x_2 

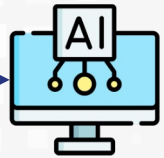
Deferral Policy

$\rho : \mathcal{X} \to \{0, 1\}$

$\rho(x_1) = 0$

$\rho(x_2) = 1$

Prediction step

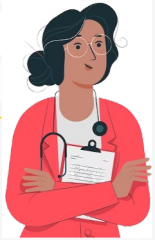


Classifier

$f : \mathcal{X} \to \mathcal{Y}$

 ↦ `Not a melanoma`

$z \in \mathcal{Z}$



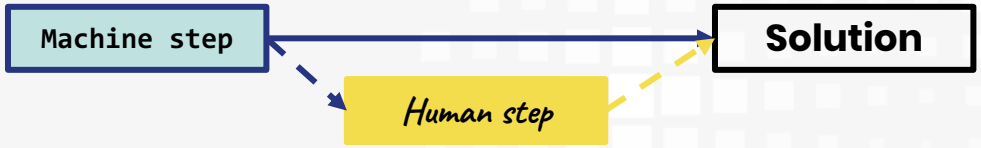
Defer to human that predicts autonomously

  ↦ `Melano ma`

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David A. Sontag. 2023. **Who Should Predict? Exact Algorithms For Learning to Defer to Humans**. ArXiv abs/2301.06197 (2023).

Learning to Abstain: pros and cons

2



Pros

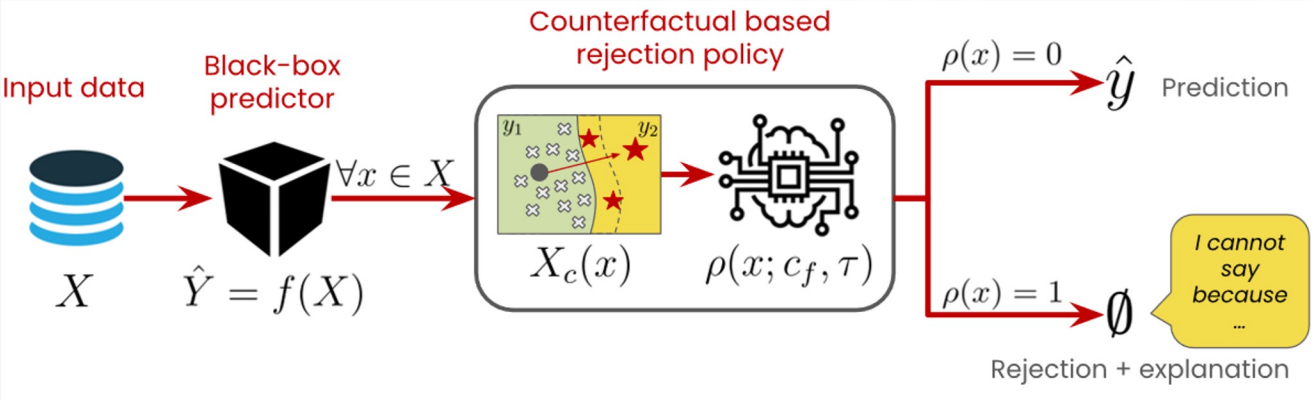
- Require human intervention only if needed
- Increase AI performance on non-rejected instances
- L2D rejects "adaptively" based on (prototypical) human predictive behavior

Cons

- No interaction with the human
- Need tons of historical human predictions
- Likely does not adapt to different humans
- **Risk of discrimination w.r.t. minority groups**
- **Opaqueness of deferral policy**

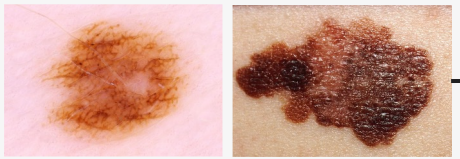
Explainability can help abstention based systems

- **Explaining the rejection of deferral policy** can help the user understand why certain instances are being redirected away from the automated decision process
- **There is little available literature on this.**



Learning together

3



Are these skin lesions melanomas?

Machine step

Yes

No

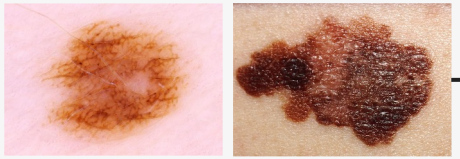
The shape is cloud-like, and the color is dark brown.

Human step

Oh no, the machine is wrong!

Learning together

3



Are these skin lesions melanomas?

Machine step

Yes

No

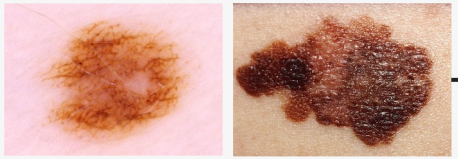
Melanomas have a round shape!

Human step

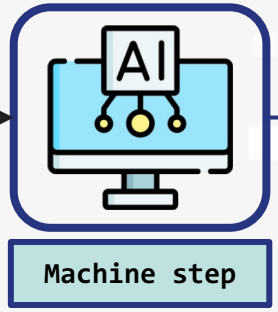
Melanomas have a round shape!

Learning together

3



Are these skin lesions melanomas?



Yes

No

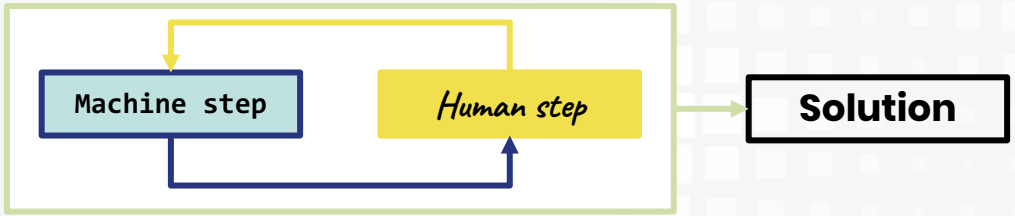
The shape is not round.

Human step

Yeee! The machine learned from me!

Learning together: pros and cons

3



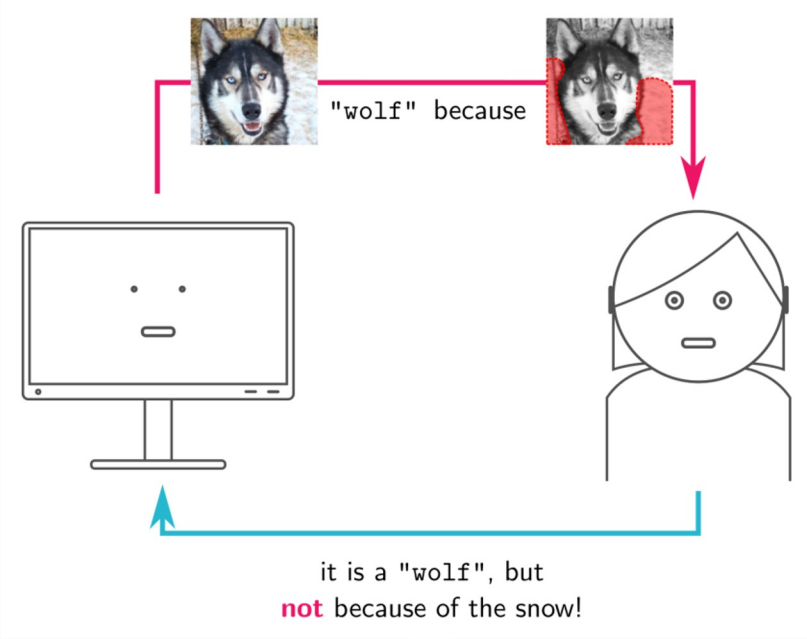
Pros

- Two-way communication
- Algorithmic correction

Cons

- The design is often task-specific, domain-specific, data-specific, language-specific, and user-specific
- Lack an abstention mechanism
- Requires initial artifacts, or artifact miner and conditioner
- No safeguards for malicious agents

Explanations as communication language



Stefano Teso - ***Toward Faithful Explanatory Interactive Machine Learning with Self-explainable Neural Nets***, Proceedings of the 3rd International Tutorial & Workshop on Interactive and Adaptive Learning (IAL'19)

FRANK

User label: $\hat{y}$

Is the record covered by a rule?

No → Are there similar records?

No → Is the model skeptical?

No → $y = \hat{y}$

Yes → Ideal Rule Check
Does $\hat{y}$ follow the rules?
Rule label: $\bar{y}$

- Yes → $y = \hat{y}$
- No → $y = \bar{y}$

Yes → Individual Fairness Check
Is $\hat{y}$ consistent with the label assigned to similar past record p?
Past label: y'_p

- Yes → $y = \hat{y}$
- No → Change the decision or relabel past ones?
  - Change → $y = y'_p$
  - Relabel → $y = \hat{y}$ Relabel with $\hat{y}$

Yes → Skeptical Learning Check
The model does not agree with $\hat{y}$. Change the decision?
Model label: $\tilde{y}$

- No → $y = \hat{y}$
- Yes → $y = \tilde{y}$
- Explain
  - Global Tree & Local Rule
  - Examples & Counterexamples *Can be relabelled*

Group Fairness Check
After every k records, the user can relabel past records to improve fairness

Takeaway message

- **Explanations** are fundamental for the future development of decision making processes where humans interact with an AI-based system.
- They can serve multiple purposes: **increase trust** between human and machine, serve as an **mean of interaction**, or can be used as a **directly interpretable** machine learning tool.

- Problem: **explanation techniques** need to be **fast and efficient.**

Thank you for your attention.