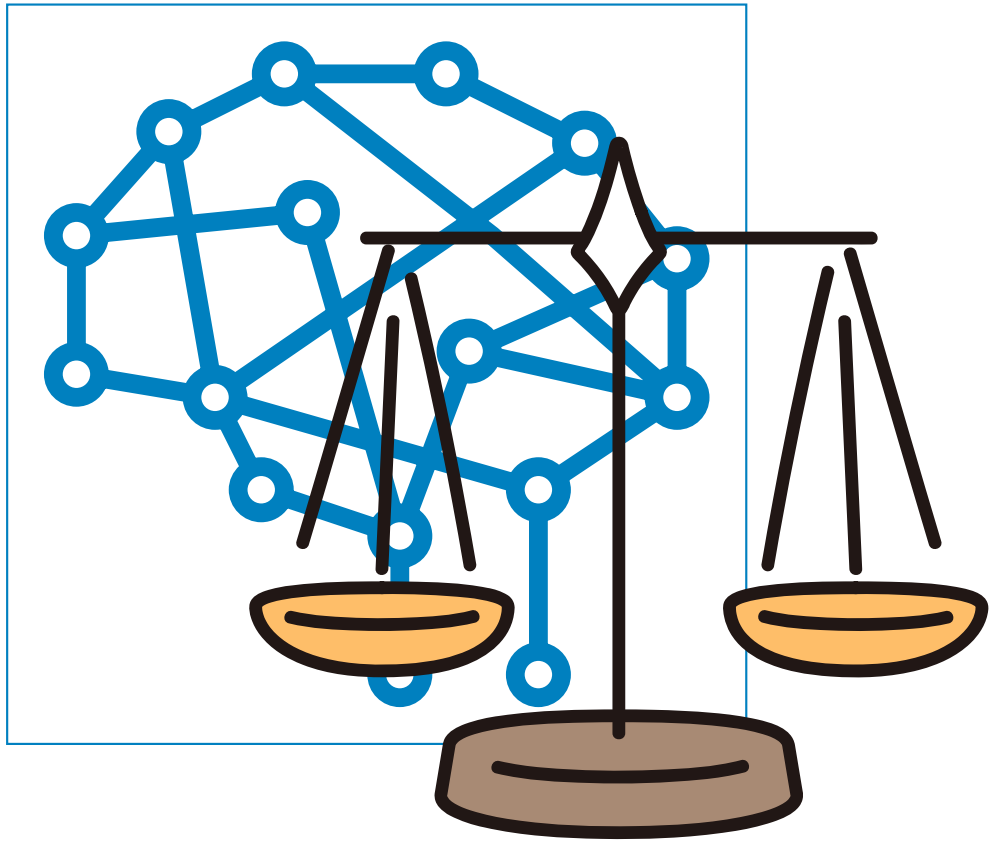
Explainable Global Fairness Verification of Tree-Based Classifiers



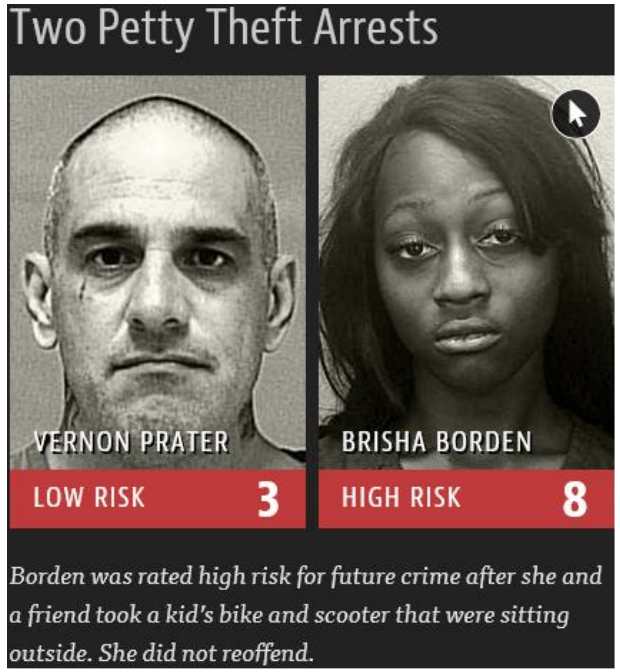
Stefano Calzavara, **Lorenzo Cazzaro**, Claudio Lucchese, Federico Marcuzzi

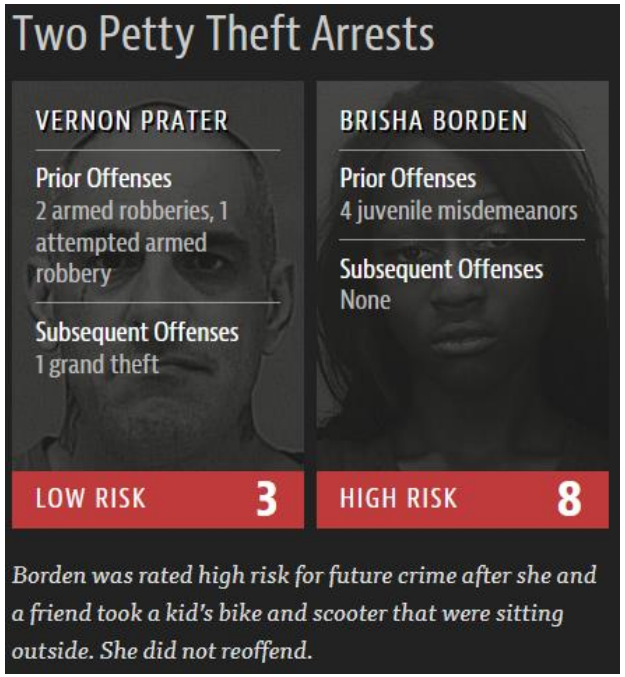
Ca' Foscari University of Venice

Accepted at the IEEE Conference on Secure and Trustworthy Machine Learning (SaTML 2023)

Is Machine Learning Fair?

Example: Machine Learning (ML) used to predict recidivity in USA*





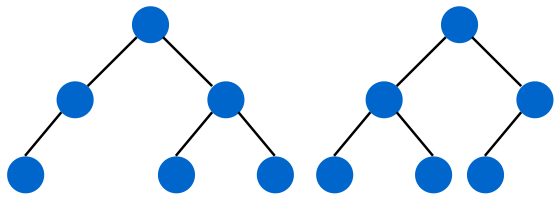
Non-recidivist black people were twice as likely to be labelled high risk than non-recidivist white people.

*https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

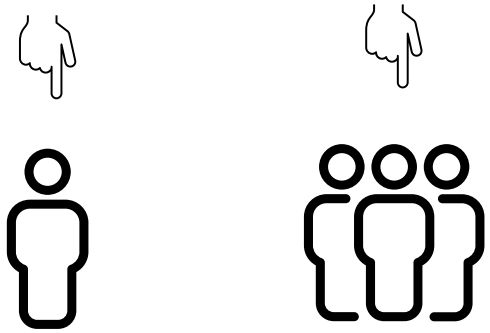
Fairness Guarantees

We need to describe the fair behaviour of a ML model by defining some **properties.**

Local Properties



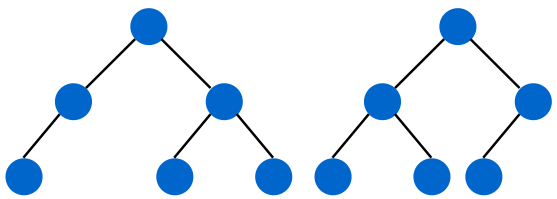
It is fair on



Test set

Global Properties

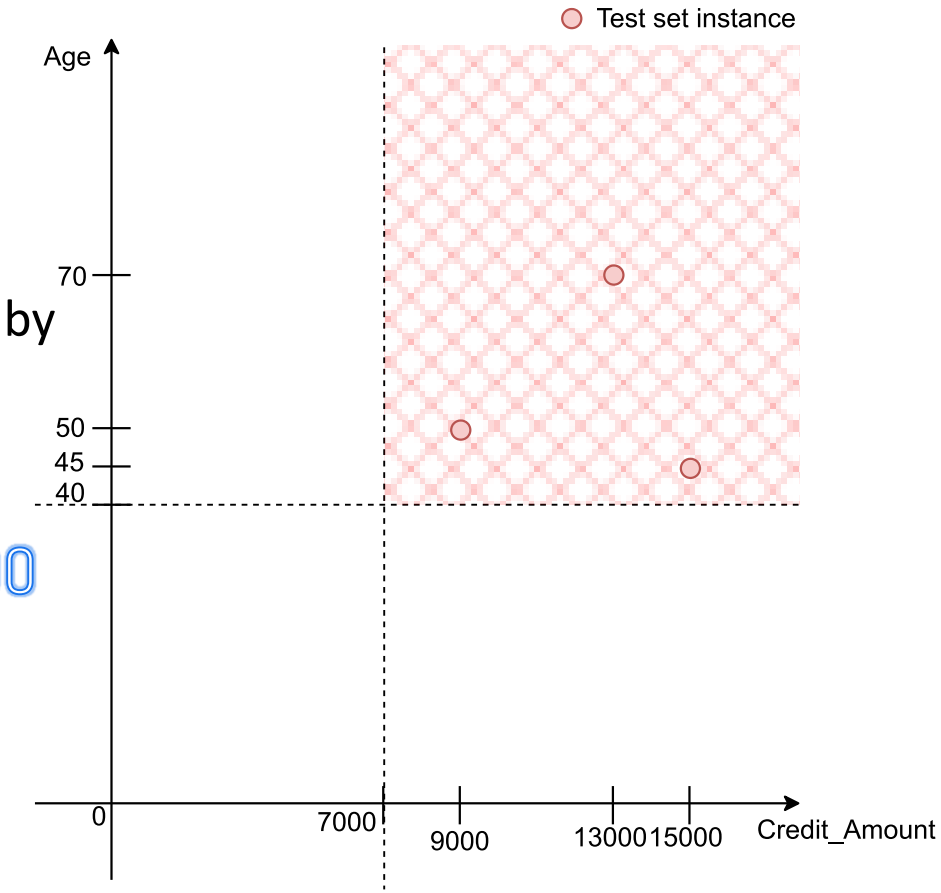
We consider *lack of causal discimination*



It is fair on people described by

Age >= 40
and
Credit_Amount >=7000

Potentially continuous and unbounded subset of instances!

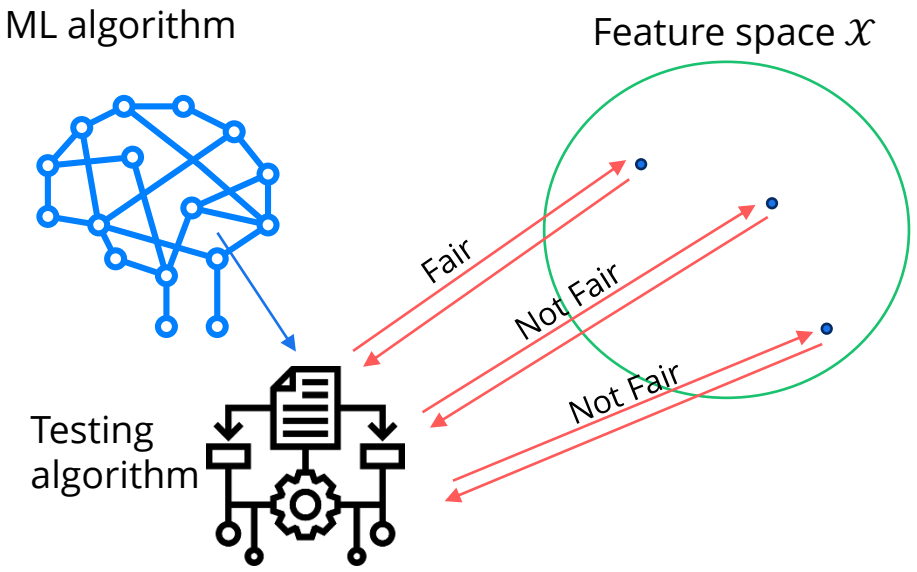


SOTA of Fairness Verification

The **explainability** of the **guarantees** is **usually neglected…**

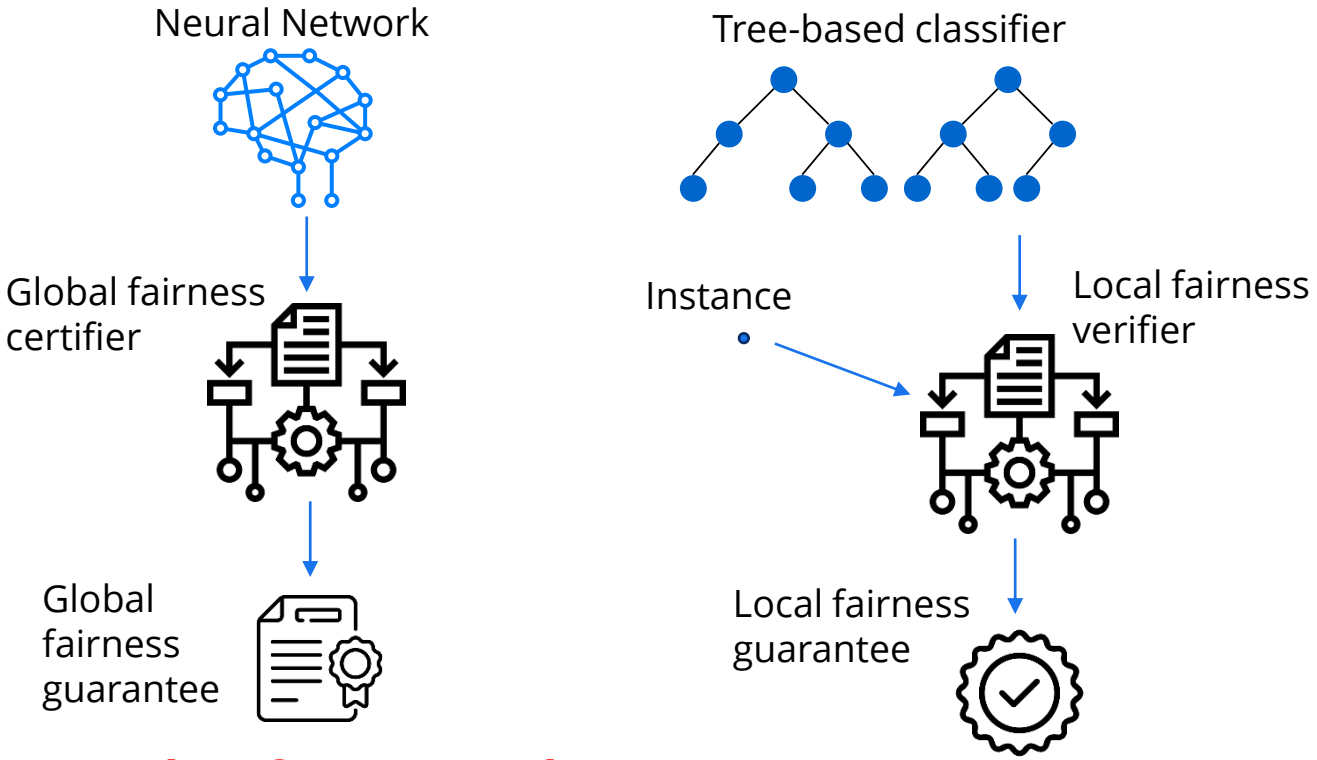
How can we prove the fairness of ML models?

Fairness Testing*[1]



Under-approximated analysis!

Formal Fairness Verification*[2-3]



Only for Neural Networks!

Supports only local properties!

*[1]A. Aggarwal, P. Lohia, S. Nagar, K. Dey, and D. Saha, *Black-box Fairness Testing of Machine Learning Models,* ESEC/SIGSOFT FSE 2019.

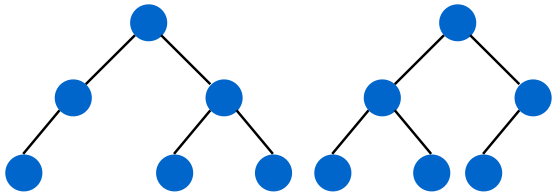
*[2]H. Khedr and Y. Shoukry, *Certifair: A Framework for Certified Global Fairness of Neural Networks,* AAAI, 2023.

*[3]F. Ranzato, C.Urban and M.Zanella, *Fairness-Aware Training of Decision Trees by Abstract Interpretation*, CIKM, 2021.

Research Problem

Problem:

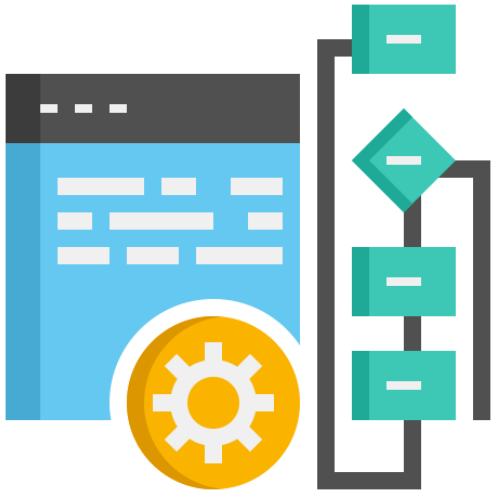
Tree based classifier T



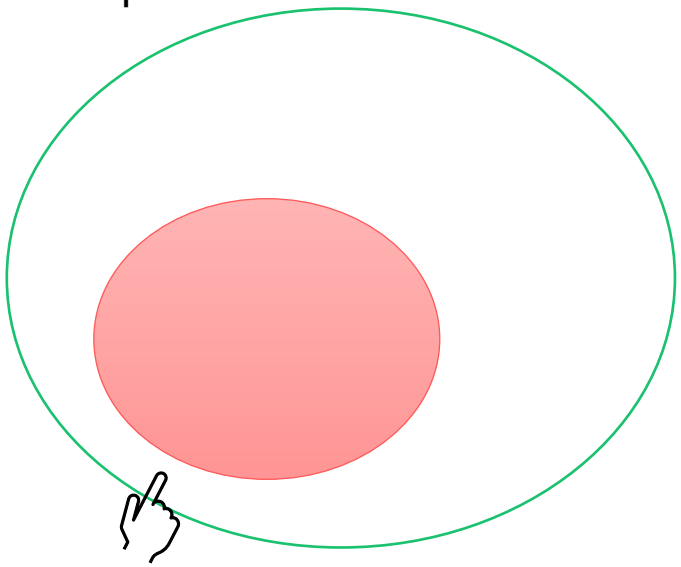
Set of sensitive features S

{sex, race}

Algorithm



Feature space $\mathcal{X}$

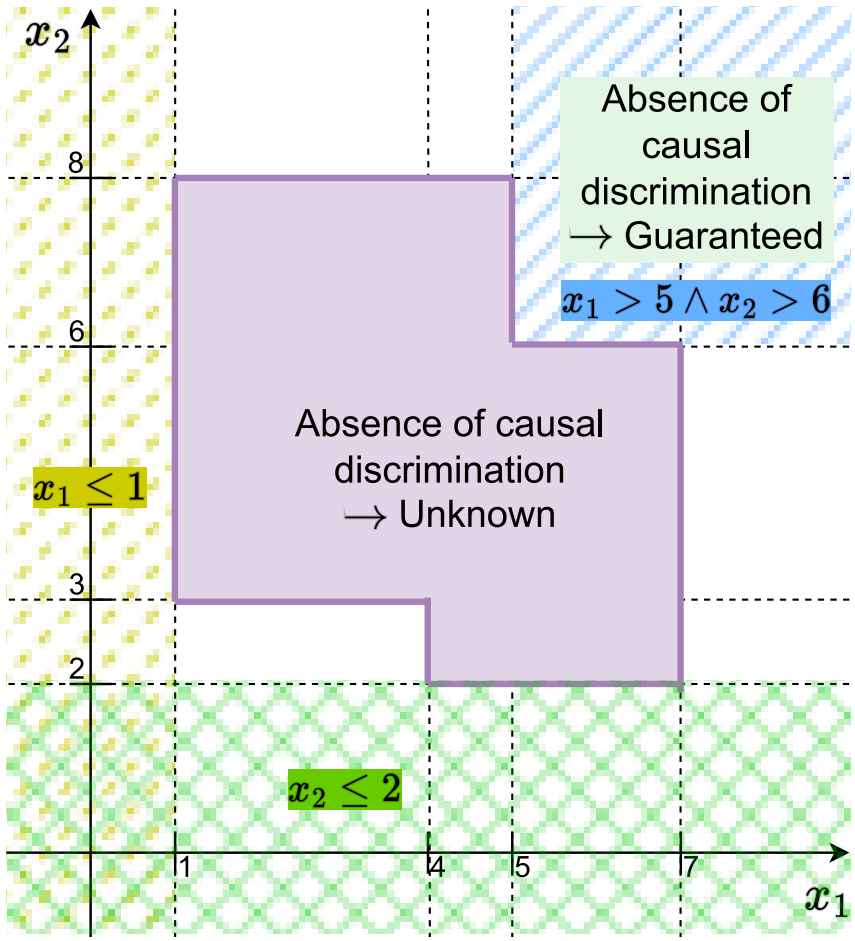


In this subset $\mathcal{X}' \subseteq \mathcal{X}$ T does not perform causal discimination w.r.t. S.
The guarantee must be *explainable*.

Contributions

Our analyzer (based on another analyzer*):

- Generates **increasingly complex sufficient conditions (logical formulas)** ensuring fairness.
- First iterations → formulas **easy to understand** (**explainable**).
- The more computational resources are available, the more complex conditions may be generated.
- We measure the **precision and the performance** of the analyzer and the **explainability of the results** of the analysis.



*S. Calzavara, L. Cazzaro, C. Lucchese, F. Marcuzzi, S. Orlando, *Beyond Robustness: Resilience Verification of Tree-Based Classifiers*, Computers&Security (2022)

Example and Conclusion

Our analysis synthesizes a set of sufficient conditions for fairness:

Conditions as **logical formulas**

Global conditions: predicate over the entire feature space

{age > 70 and job = «prof»,
credit_account < 4000 and age < 35 and housing = «rent»}

Explainable formulas: readily understandable

Our analysis is precise, explainable, reasonably efficient and proved sound and complete (details in the full paper)!

Lorenzo Cazzaro
Ph.D. student in Computer Science

@LorenzoCazz

lorenzo.cazzaro@unive.it

Ca' Foscari University of Venice

Lorenzo Cazzaro

LorenzoCazzaro

https://lorenzocazzaro.github.io/

Thank you! Questions?