Trustworthy Machine Learning Through the Lens of Combinatorial Optimization and Operations Research

Keywords: Fairness, Interpretability, Explainability, Privacy

Mohamed Siala, Marie-José-Huguet, Emmanuel Hebrard

ROC Research Team, DO Department

Context

- Machine learning now permeates daily life and high-stakes decisions, making **trustworthiness** more critical than ever.
- Trust spans multiple dimensions: **philosophy**, **bias mitigation**, and **technical safety**.
- Advances in **combinatorial optimization** and **operations research** open new paths to address trust-related challenges.

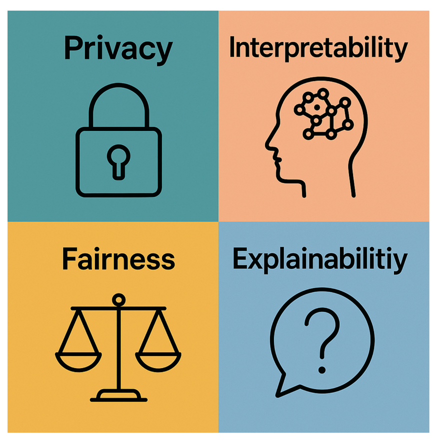

Figure 1: Four Facets of Trust Explored in This Work

Contributions

- Fair rule lists using backtracking search and mixed-integer linear programming [1, 2]
- Optimal decision trees and diagrams using Boolean satisfiability, backtracking search, and dynamic programming [3, 4]
- Dataset reconstruction from interpretable models [5]
- Trustworthy explanations using Boolean satisfiability [6]
- Generalizing statistical fairness to unseen data using sample-robust optimization [1]
- Sensitive attribute reconstruction using constraint programming [7]

Example of Interpretable Models

- A *rule list* model provides interpretable predictions by applying an ordered sequence of if–then rules to the input data.
- *Default of Credit Card Clients* dataset: each individual is represented by demographic information, credit data, and payment history. The goal is to predict whether a person will default on their credit card payment.

  1. If `Delay` $=2$ months and `Age` >60, then **predict Default**.
  2. Else if `Education` $=$ University and `Sex` $=$ Male, then **predict No Default**.
  3. Else if `Marriage` $=$ Single and `Bill`$>50k$, then **predict Default**.
  4. Else **predict No Default**.

  Figure 4: Example of a rule list for the *Default of Credit Card Clients* dataset.

- The model above might exhibit a gender bias. Without algorithmic considerations for fairness, unbiased outcomes cannot be guaranteed.

Example: Learning Fair Rule Lists

- There exist multiple fairness measures. For a statistical measure $\mathcal{M}$, we aim to ensure that for two protected groups $\mathcal{A}, \mathcal{B}$ (e.g., male and female),

$$|\mathcal{M}(\mathcal{A}) - \mathcal{M}(\mathcal{B})| < \epsilon.$$

- Let X be the training data and Y their true labels. For a given fairness measure $\mathcal{M}$, the fair rule list problem seeks a rule list that minimises prediction error subject to fairness constraints. Let $\mathcal{R}$ be the set of all possible rules:

$$r^* = \arg\min_{r \in \mathcal{R}} \texttt{error}(r, X, Y)$$
$$\text{s.t.} \quad \texttt{fairness}(r, X, Y) \leq \epsilon.$$

- We solve this problem using a backtracking search. At each node, an integer linear program (ILP) ensures that the joint misclassification–fairness constraint remains feasible.

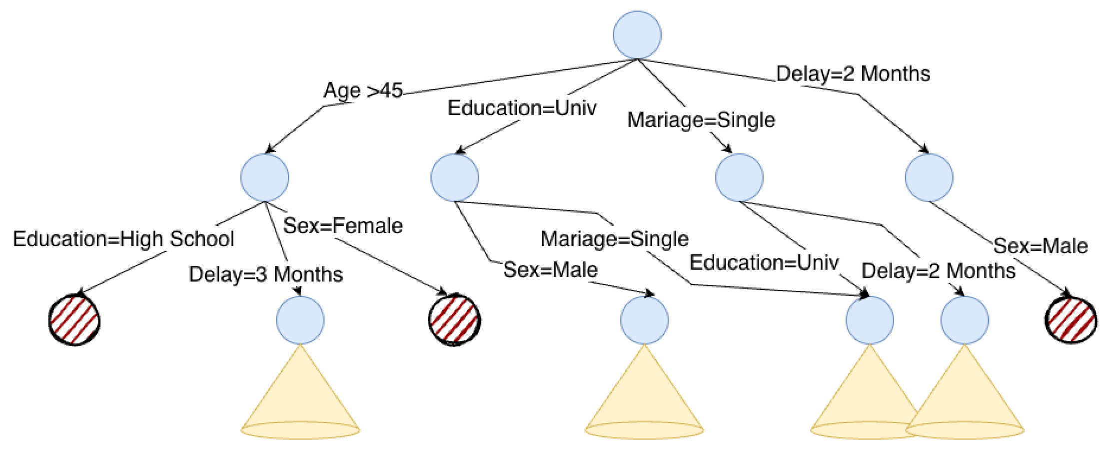

Figure 2: Pruning the search space with ILP. Each node represents a set of candidate models; dashed nodes are pruned by the ILP model.

- We also address the challenging problem of *generalizing fairness*, using mixed-integer programming within a sample-robust optimization framework to guarantee fairness under uncertainty.
- `FairCORELS` is a **multi-objective tool** designed to learn fair rule lists.

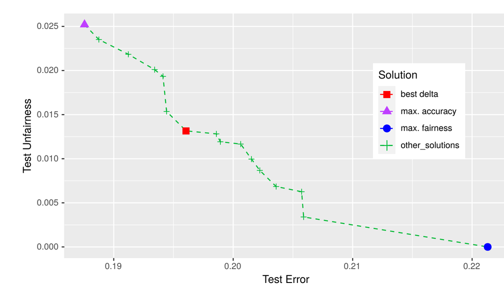

Figure 3: Example trade-offs produced by `FairCORELS` for the Statistical Parity fairness metric on the *Default of Credit Card* dataset.

References

[1] Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Improving fairness generalization through a sample-robust optimization method. *Mach. Learn.*, 2023.

[2] Ulrich Aïvodji, Julien Ferry, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Leveraging integer linear programming to learn optimal fair rule lists. In *CPAIOR*, 2022.

[3] Emir Demirovic, Anna Lukina, Emmanuel Hebrard, Jeffrey Chan, James Bailey, Christopher Leckie, Kotagiri Ramamohanarao, and Peter J. Stuckey. Murtree: Optimal decision trees via dynamic programming and search. *J. Mach. Learn. Res.*, 2022.

[4] Hao Hu, Marie-José Huguet, and Mohamed Siala. Optimizing binary decision diagrams with maxsat for classification. In *AAAI*, 2022.

[5] Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Probabilistic dataset reconstruction from interpretable models. In *IEEE SaTML*, 2024.

[6] Mohamed Siala, Jordi Planes, and João Marques-Silva. On trustworthy rule-based models and explanations. In *ECML PKDD*, 2025.

[7] Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Exploiting fairness to enhance sensitive attributes reconstruction. In *IEEE SaTML*. IEEE, 2023.