

# Fairness in classification algorithms

Conformal Prediction and Optimal Transport

---

**Rubén Armañanzas** and **Jesús López Fidalgo**

Alberto García Galindo, Elena Martín de Diego

Time to Share: Ética e Inteligencia Artificial

TECNUN, San Sebastián

December 18, 2023



Universidad  
de Navarra

DATAI  
INSTITUTE OF DATA SCIENCE  
AND ARTIFICIAL INTELLIGENCE

# OUTLINE.



Universidad  
de Navarra

DATAI  
INSTITUTE OF DATA SCIENCE  
AND ARTIFICIAL INTELLIGENCE

Introduction

**01**

Fairness in Machine Learning

**02**

Conformal Prediction


**03**

Optimal Transport

Conclusions

# INTRODUCTION.

Loan approval  
Candidate screening  
Clinical diagnosis



Decision-making based on algorithms can have a **substantial impact** in our lives.

Despite promise, the application of machine learning may have **unintended consequences**.

## Criminal justice: recidivism algorithms (COMPAS)

- Predicting if a defendant should be imprisoned.

ProPublica Analysis of COMPAS algorithm (Angwin et al., 2022)

	White	Black
Wrongly labelled as high-risk	23.5%	44.9%
Wrongly labelled as low-risk	47.7%	28.0%



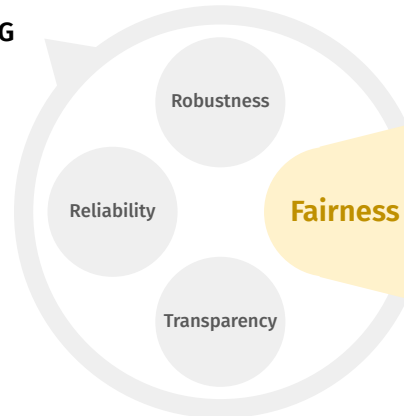
## INTRODUCTION.

Decision-making based on algorithms can have a **substantial impact** in our lives.

Despite promise, the application of machine learning may have **unintended consequences**.

As the integration of data-driven algorithms into safety-critical systems has become more widespread, so has the **ethical concerns** about its misuse.

### TRUSTWORTHY MACHINE LEARNING



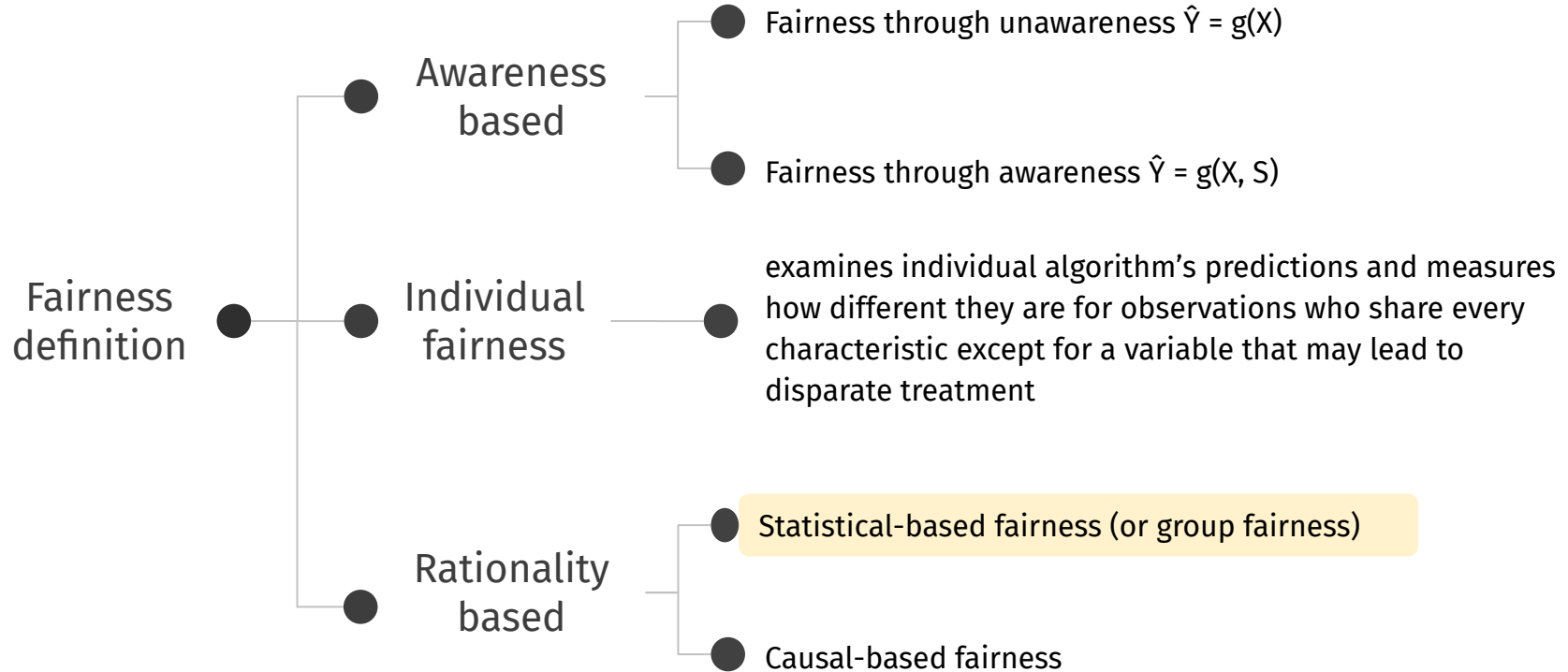
#### Fairness

Ensure that **predictive algorithms do not discriminate** towards any individual or subgroups with respect to their sensitive attributes.

*Age, gender, ethnicity, disability, sexual orientation...*



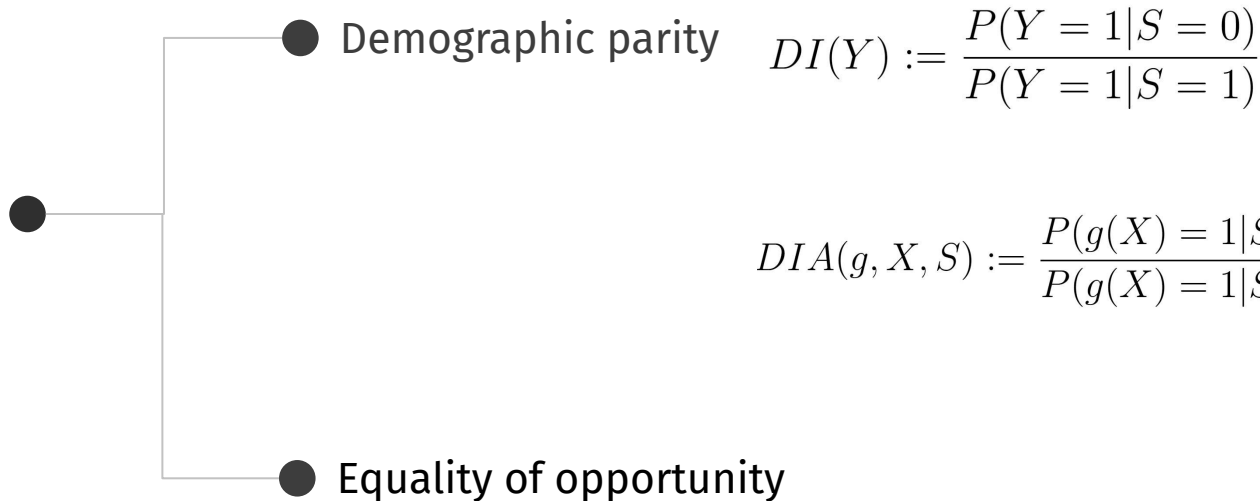
## FAIRNESS DEFINITIONS.



# FAIRNESS DEFINITIONS.



Statistical-based  
fairness



$$P(g(X) = 1|S = 0, Y = 1) = P(g(X) = 1|S = 1, Y = 1)$$

## FAIR CLASSIFICATION.

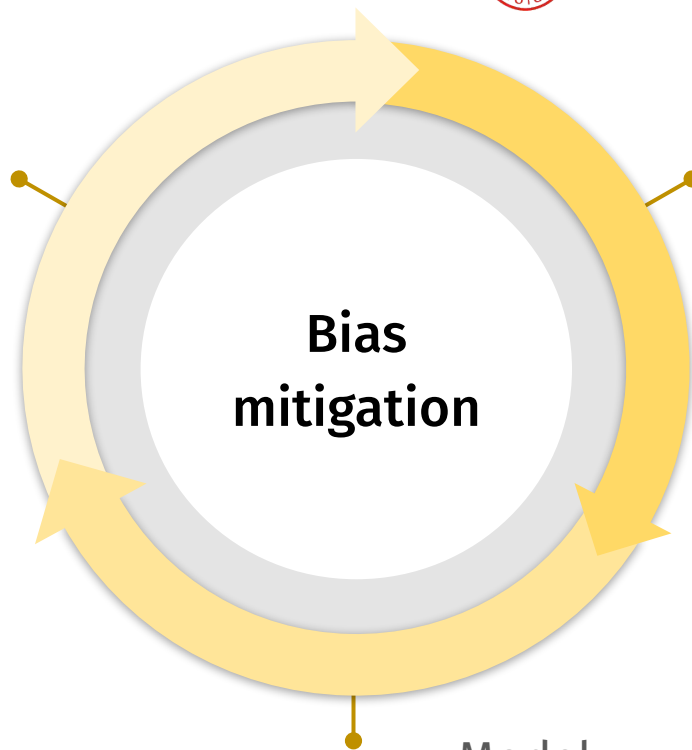


Universidad  
de Navarra

DATAI  
INSTITUTE OF DATA SCIENCE  
AND ARTIFICIAL INTELLIGENCE

Outcome  
Post-processing

Data  
Pre-processing



Model  
In-processing



Introduction

**01**

Fairness in Machine Learning

**02**

Conformal Prediction

**03**

Optimal Transport

Conclusions





# CONFORMAL PREDICTION.

## Conformal Prediction

Quantify uncertainty through *prediction sets with guarantees*

*Main idea:*

**Calibrate** a trained machine learning model  with an external calibration dataset.

**Example: Automated diagnosis of COVID19 (Angelopoulos et al, 2022)**

Underlying deep learning model: ResNet-18.

**Prediction sets** are computed with **coverage** guarantees: **contain the true diagnosis with 90% of confidence.**

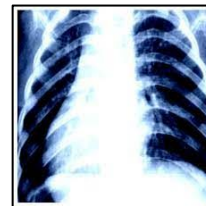


[**covid19**]

**Less uncertainty** (shorter interval)



[**bacterial, covid19**]



[**bacterial, covid19, normal**]

**More uncertainty** (larger interval)

# CONFORMAL PREDICTION AND FAIRNESS.



Universidad  
de Navarra

DATAI  
INSTITUTE OF DATA SCIENCE  
AND ARTIFICIAL INTELLIGENCE

Convey predictions with uncertainty is a **fundamental way to support fair decision-making.**

$$f_1 = \frac{1}{n} \sum_{i=1}^n |\mathcal{C}(x_i)|$$

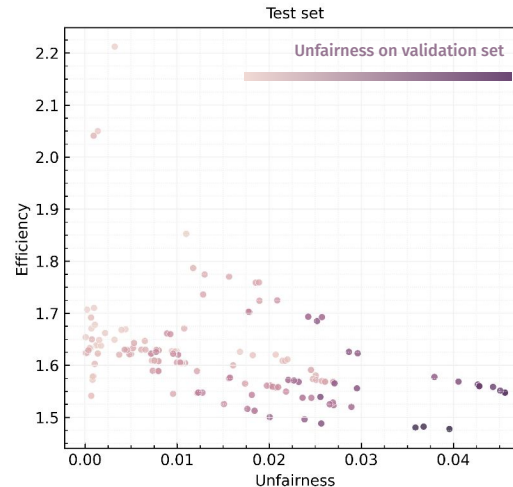
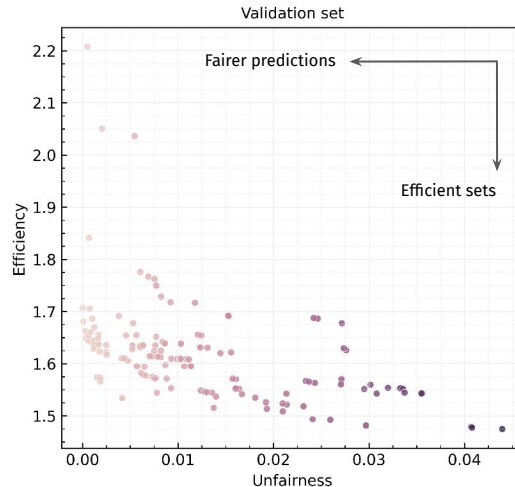
*Objective*

**Compute Efficient Prediction Sets with Fairness Guarantees**

$$f_2 = |\text{Cov}_{\text{male}} - \text{Cov}_{\text{female}}|$$

*Proposal*

**Tune the hyperparameters of the underlying model with multiobjective evolutionary algorithms**



Dataset: **Adult Income**

Prediction: **Gross annual income (3 tiers)**

Sensitive attribute: **gender**

Confidence level: **90%**

Optimized classifier: **decision tree**

# OUTLINE.



Universidad  
de Navarra

DATAI  
INSTITUTE OF DATA SCIENCE  
AND ARTIFICIAL INTELLIGENCE

Introduction

**01**

Fairness in Machine Learning

**02**

Conformal Prediction

**03**

Optimal Transport

Conclusions

# OPTIMAL TRANSPORT.



Universidad  
de Navarra

DATAI  
INSTITUTE OF DATA SCIENCE  
AND ARTIFICIAL INTELLIGENCE

**Discrimination** of the classification procedures appears as soon as the prediction and the protected attribute are too closely related.

→ **Goal:** achieve statistical parity  $\mathbb{P}(g(X) = 1 \mid S = 0) = \mathbb{P}(g(X) = 1 \mid S = 1)$

→ **Mathematical formalization of bias**  $\mathcal{L}(X \mid S = 0) \neq \mathcal{L}(X \mid S = 1)$

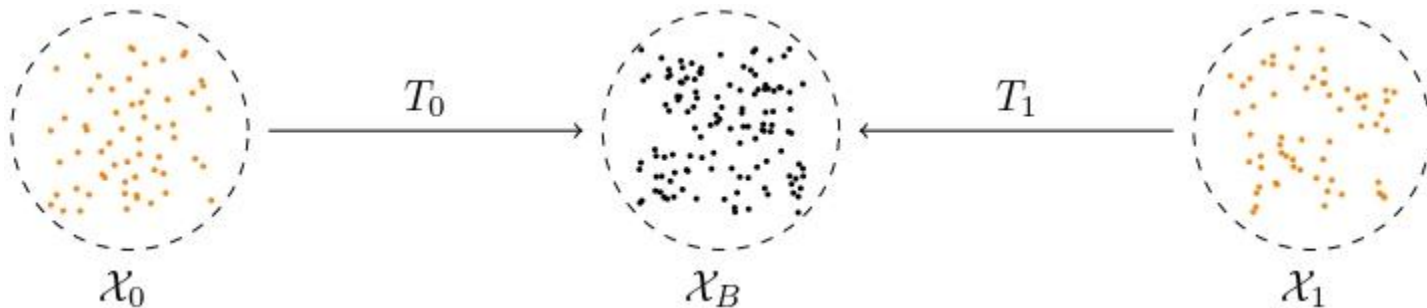
→ **Methodology**  $\mathcal{L}(\tilde{X} \mid S = 0) = \mathcal{L}(\tilde{X} \mid S = 1)$

# OPTIMAL TRANSPORT.



Methodology  $\mathcal{L}(\tilde{X} | S = 0) = \mathcal{L}(\tilde{X} | S = 1)$

$\mathcal{L}(g(\tilde{X}) | S = 0) = \mathcal{L}(g(\tilde{X}) | S = 1)$



$\mu_0 := \mathcal{L}(X|S = 0)$      $\mu_B \in \operatorname{argmin}_{\nu \in \mathcal{P}_2} \{ \pi_0 \mathcal{W}_2^2(\mu_0, \nu) + \pi_1 \mathcal{W}_2^2(\mu_1, \nu) \}$      $\mu_1 := \mathcal{L}(X|S = 1)$

# OUTLINE.



Universidad  
de Navarra

DATAI  
INSTITUTE OF DATA SCIENCE  
AND ARTIFICIAL INTELLIGENCE

Introduction

**01**

Fairness in Machine Learning

**02**

Conformal Prediction

**03**

Optimal Transport

Conclusions

# CONCLUSIONS AND FUTURE WORK.



Universidad  
de Navarra

DATAI  
INSTITUTE OF DATA SCIENCE  
AND ARTIFICIAL INTELLIGENCE

Machine learning have penetrated **safety-critical domains**.

There is a dire need to develop **technical solutions** for addressing the problem of unfair decisions.



- Novel **calibration** procedures
- Fair classifiers with **reject option**
- Conformal Risk Control with **Fairness Guarantees**



- Extended total repair to online scenarios
- Fairness in **Large Language Models**
- Fairness in different **data structures**
- Ethical **guidelines** for auditing AI models

We need the ethical fundamentals to do all this!

# Fairness in classification algorithms

Conformal Prediction and Optimal Transport

---



Universidad  
de Navarra

DATAI  
INSTITUTE OF DATA SCIENCE  
AND ARTIFICIAL INTELLIGENCE

**Thank you for your attention!**