Bias and Fairness in AI-Based Systems

Giordano d'Aloisio

SWEN / Software Engineering Research Group

Università degli Studi dell'Aquila, Italy

SWEN research group

UNIVERSITÀ DEGLI STUDI DELL'AQUILA

About Me

Giordano d'Aloisio

I'm a Postdoctoral Researcher at the University of L'Aquila in Italy where I obtained the European PhD in Information and Communication Technology in 2025 under the supervision of Prof. Antinisca Di Marco.

From Febraury to August 2024 I was a visiting researcher at the University College London (UCL) as a member of the SOLAR research group working with Prof. Federica Sarro.

From 2019 to 2024, I was a member of the Territori Aperti project where I was responsible of the Data Integration activity.

My research is mostly focused on quality aspects of learning-based systems, with a particular attention on software bias and fairness.



University of L'Aquila
DISIM Department
via Vetoio
67100 L'Aquila, Italy

Research Interests

- Software Fairness
- Software Sustainability
- Empirical Software Engineering
- Search-Based Software Engineering
- Human Aspects in Computer Science

Where I'm From



L'AQUILA
2026
Capitale italiana della Cultura



Index

Key Concepts to Bias and Fairness

Bias Assessment

Bias Mitigation

Live example

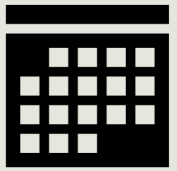
https://colab.research.google.com/drive/1rES8llx70jmH4oW2UVWj6yKDbO8BDT6u?usp=sharing

Key Concepts to Bias and Fairness

Supervised Machine Learning

Training

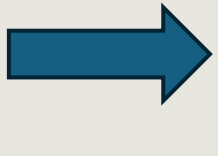
Dataset



Features



Label



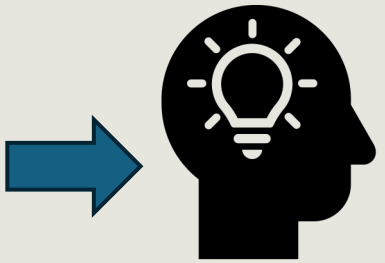


Machine Learning Model

Testing



Unseen Features



Trained ML Model



Prediction

Face Recognition System









Trained ML Model



Face

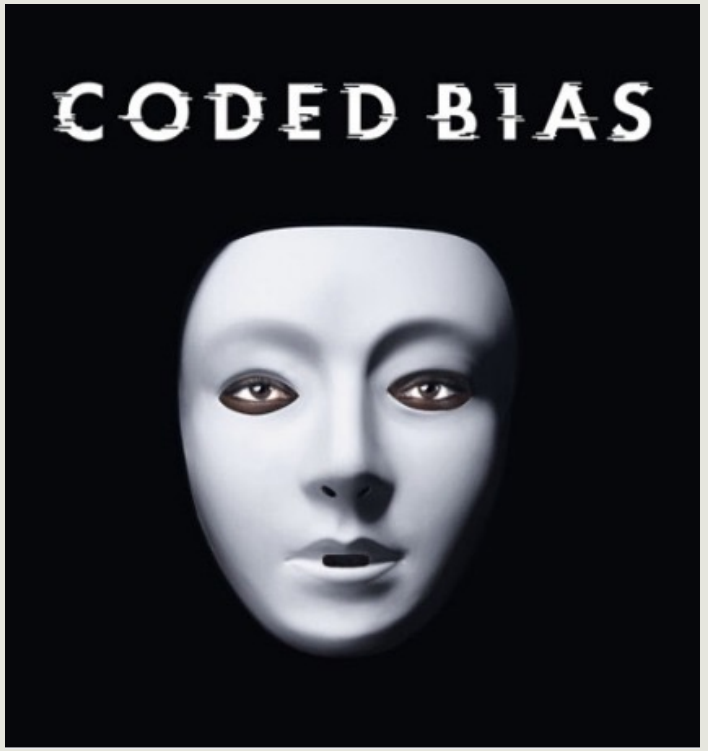


Not a face

Coded Bias

- In 2018, a researcher at MIT was working with Amazon's face recognition systems
- The system was not able to recognise her face
- But when she wore a white mask, the system recognised her

The system was unable to recognise black women

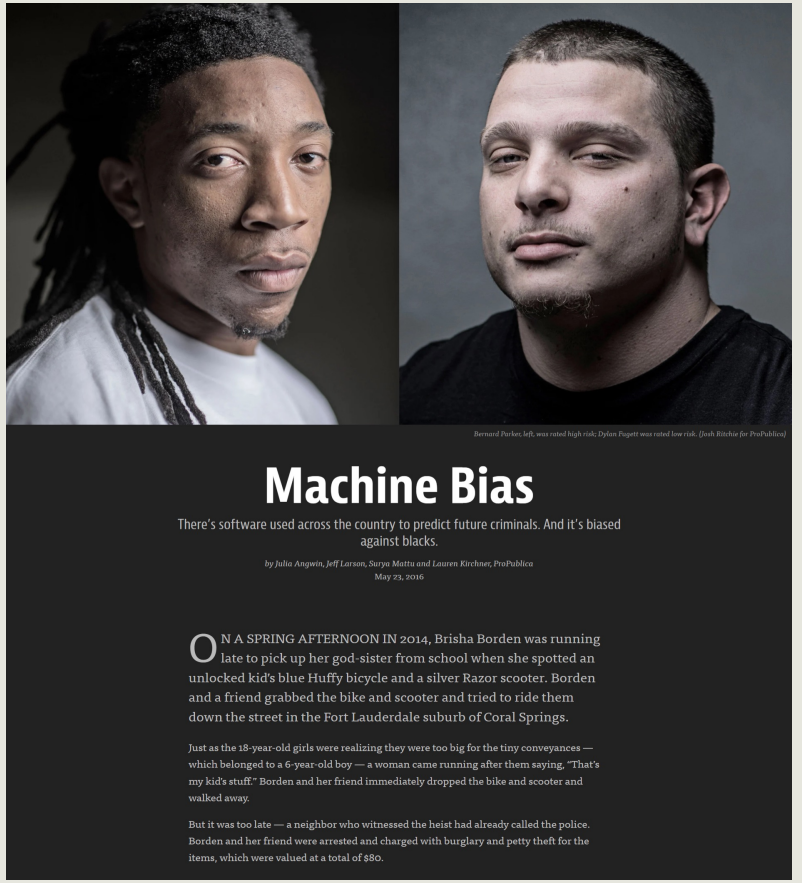


Another example

- **COMPAS** was an ML algorithm used by some courts in the US to predict recidivism of condemned people

- A study showed that, given two people with the same features but different race, the system was giving higher probability of recidivism to black people

The system was systematically favouring white people



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

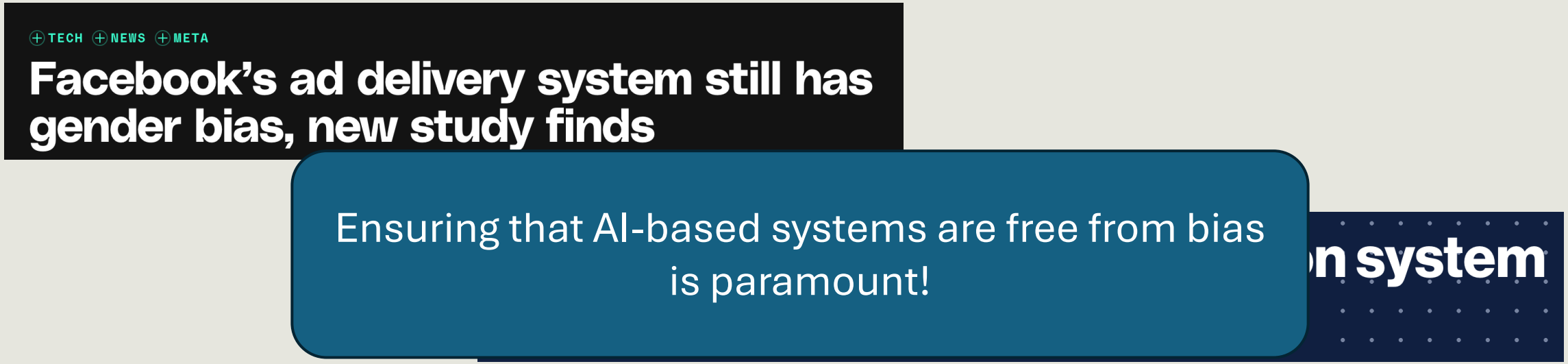
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Other examples…



A-level students launch legal challenge over results chaos as pressure grows on Gavin Williamson to resign

A crowd-funding campaign by the Good Law Project on behalf of six students has raised £65,000

Definitions

- **Bias** is defined as: “the systematic discrimination or favouritism of an AI-based system towards individuals or groups based on their inherent characteristics” [1]

- In other words, **bias** is a deviation from expected data patterns. This deviation implies favouritism or discrimination in *predictions* toward a *specific group* identified by a set of *sensitive variables*



[1] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Key Concepts

- **Sensitive variables:** attributes of the dataset that are socio-culturally critical for the application of AI-based models (e.g., *gender, ethnicity, age, religion, etc.*)

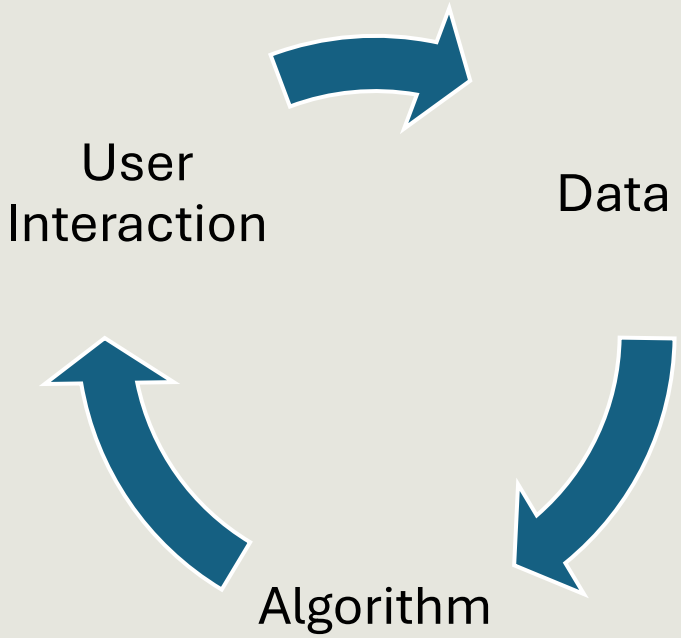
- **Positive outcome:** a specific system outcome that may lead to discrimination (e.g., *granting a loan*, *admission to a university*, *non-recidivism*, etc.)

- **Privileged and unprivileged groups:** items identified by one or more sensitive attributes that are systematically privileged/unprivileged by the system (e.g., *men vs women, Caucasian vs Non-Caucasian, etc.*)

Sources of Bias

In general, bias can originate:

- From the **dataset**
- From the **algorithm**
- From the **people** who use a system



Bias Assessment

Fairness Definitions

- **Algorithmic Fairness** (or simply **Fairness**) can be defined as "the absence of any prejudice or favouritism towards an individual or a group."

- The group is usually identified by a set of *sensitive* (or protected) *variables*.

- There are more than 23 fairness definitions from the literature, classified into two categories:

  - **Individual Fairness:** give similar predictions to similar individuals
  - **Group Fairness:** treat different sub-groups equally

Fairness Definitions: Statistical/Demographic Parity

The probability of an item being positively classified should be the same regardless of the value(s) of the sensitive variable(s):

$$P\left(\hat{Y} = 1 \middle| S = 0\right) = P(\hat{Y} = 1 | S = 1)$$

Corresponding Metric

- **Statistical Parity Difference**

$$P\left(\hat{Y} = 1 \middle| S = 0\right) - P(\hat{Y} = 1 | S = 1)$$

An absolute value ≤ 0.2 means fairness

- **Disparate Impact**

$$\frac{P(\hat{Y} = 1 | S = 0)}{P(\hat{Y} = 1 | S = 1)}$$

A score between 0.8 and 1.2 means fairness

Example

$$P(\hat{Y} = 1 | S = 0) = P(\hat{Y} = 1 | S = 1)$$

Demographic parity

20 applicants (50% from **Group A**)
14 approvals (50% from **Group A**)

GROUP A

TRUE \ PREDICTED	Deny	Approve
Deny	1	2
Approve	2	5

GROUP B

TRUE \ PREDICTED	Deny	Approve
Deny	2	4
Approve	1	3

Fairness Definitions: Equal Opportunity

The probability of having a positive prediction from a system conditioned to a positive value of the ground-truth label should be equal, regardless of the value of the sensitive variable(s):

$$P\left(\hat{Y} = 1 \middle| S = 0, Y = 1\right) = P(\hat{Y} = 1 | S = 1, Y = 1)$$

Corresponding Metric

Equal Opportunity Difference

$$\left(\frac{TP}{TP + FN}\right)_{unpriv} - \left(\frac{TP}{TP + FN}\right)_{priv}$$

An absolute value ≤ 0.15 means fairness

Example

$$\left(\frac{TP}{TP + FN}\right)_{unpriv} - \left(\frac{TP}{TP + FN}\right)_{priv}$$

Equal opportunity

Group A: 66% true positive rate: 4/(4+2)
Group B: 66% true positive rate: 2/(1+2)

GROUP A

TRUE \ PREDICTED	Deny	Approve
Deny	3	1
Approve	2	4

GROUP B

TRUE \ PREDICTED	Deny	Approve
Deny	6	1
Approve	1	2

Fairness Definitions: Equal Accuracy

The accuracy of the system should be equal despite the value of the sensitive variable(s):

$$P(\hat{Y} = y | S = 0, Y = y) = P(\hat{Y} = y | S = 1, Y = y) \quad y \in \{0,1\}$$

Corresponding metric

Equal Accuracy Difference

$$\left(\frac{TP+TN}{TP+TN+FP+FN}\right)_{unpriv} - \left(\frac{TP+TN}{TP+TN+FP+FN}\right)_{priv}$$

An absolute value ≤ 0.15 means fairness

Example

$$\left(\frac{TP+TN}{TP+TN+FP+FN}\right)_{unpriv} - \left(\frac{TP+TN}{TP+TN+FP+FN}\right)_{priv}$$

Equal accuracy

Group A: 80% accurate: (6+2)/10
Group B: 80% accurate: (4+4)/10

GROUP A

TRUE \ PREDICTED	Deny	Approve
Deny	2	1
Approve	1	6

GROUP B

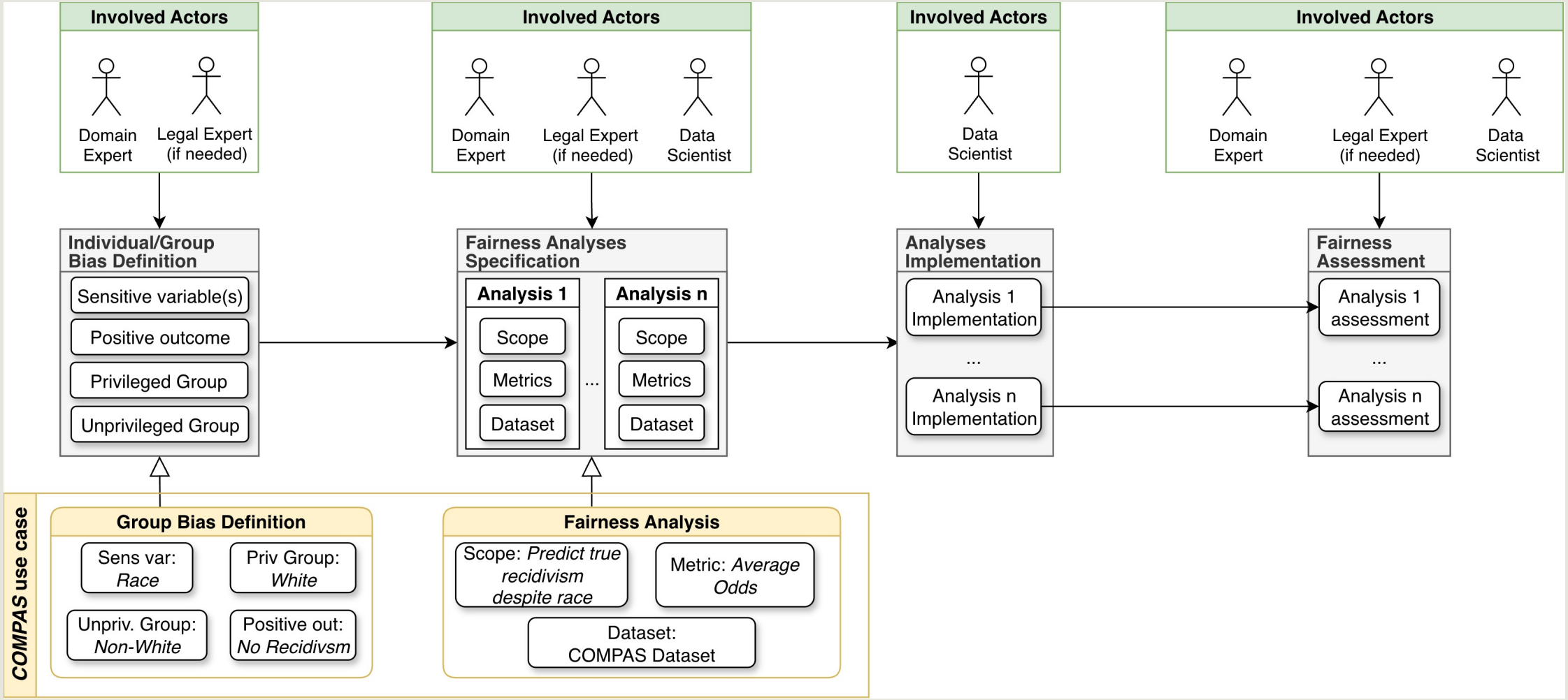
TRUE \ PREDICTED	Deny	Approve
Deny	4	2
Approve	0	4

Which Definition to Adopt?

The choice of the fairness definition to adopt is domain-dependent:

- Do you need to achieve perfect parity?
  - Use Demographic Parity
- Do you care mostly about positive predictions?
  - Use Equal Opportunity
- Do you care about the overall model accuracy?
  - Use Equal Accuracy

A General Fairness Assessment Process [2]



[2] d'Aloisio, G., Di Sipio, C., Di Marco, A., & Di Ruscio, D. (2025). How fair are we? From conceptualization to automated assessment of fairness definitions. *Software and Systems Modeling*, 1-27.

Bias Mitigation

Bias Mitigation

- Bias mitigation is the process of removing bias from a dataset or a model
- Removing the sensitive features from the training set is not enough to mitigate bias. Indeed, it may even decrease the model performance
- For this reason, the literature has proposed multiple **bias mitigation methods**

Bias Mitigation Methods

- 341 bias mitigation methods have been identified in the literature [3]
- Approaches are divided into three categories:
  - **Pre-processing:** transform the training data so that the underlying discrimination is removed.
  - **In-processing:** change ML algorithms' behaviour to remove discrimination during training.
  - **Post-processing:** remove the learned bias from an already trained ML model

[3] Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM J. Responsib. Comput.*, *1*(2)

Pre-processing Techniques

- Change the input data to remove intrinsic bias
- 132 pre-processing methods have been identified
- Main strategies:
  - **Relabeling and Perturbation:** change the value of the features/label for some instances
  - **Sampling:** change the data distribution
  - **Feature Engineering:** replace biased features with new features that are unbiased
  - **Representation Learning:** learn fair latent representation

In-processing Techniques

- Change the training process of an ML model to make it fair
- 212 in-processing methods proposed in the literature
- Main strategies:
  - **Regularisation and Constraints:** add additional constraints to the learning process
  - **Adversarial Learning:** train an adversarial model that maximises fairness together with the main ML model
  - **Compositional Learning:** train multiple models for different subgroups
  - **Adjusted Learning:** change the training behaviour or design new algorithms

Post-processing Techniques

- Work on an already trained ML model to remove the learned bias
- 56 post-processing methods proposed in the literature
- Main strategies:
  - **Input Correction:** change the input features to remove any sensitive information
  - **Classifier Correction:** re-train an already trained classifier to make it fairer
  - **Output Correction:** modify the predictions of a biased classifier

Fairness vs Accuracy

- Fairness improvement always implies changing the model's training/prediction behaviour
- Therefore, it has an impact on the model's performance
- When designing a bias mitigation algorithm, researchers should always consider the fairness – accuracy trade-off

Pre- vs In- vs Post-Processing

Pre-Processing

Pro

- (Mainly) model agnostic
- Easy to reuse

Const

- May mostly impact model accuracy
- Hard to guarantee fairness under distribution shift

In-Processing

Pro

- Strong theoretical guarantee
- Optimisation control

Const

- (Generally) Model-specific
- Requires changing the internal model behaviour

Post-Processing

Pro

- Black-box compatible
- No retraining needed (popular with LLMs)

Const

- Limited effectiveness
- May impact accuracy

Which category to choose?

- **General rule:** the sooner you can apply a technique, the better
- Can you change the training data?
  - Use **pre-processing** methods
- Can you train the model?
  - Use **in-processing** methods
- Do you have to mitigate bias of an already trained model?
  - Use **post-processing** methods

Live example



https://colab.research.google.com/drive/1rES8llx70jmH4oW2UVWj6yKDbO8BDT6u?usp=sharing

Thank you for your attention!

giordano.daloisio@univaq.it