ENFORCING FAIRNESS VIA CONSTRAINT INJECTION WITH FAUCI

2nd Aequitas Workshop on Fairness and Bias in AI at ECAI 2024

🎤 *Matteo Magnini*, Giovanni Ciatto, Roberta Calegari, Andrea Omicini

📧 matteo.magnini@unibo.it

CONTEXT

WHAT DO WE MEAN BY FAIRNESS?

Fairness has different meanings to us depending on our *personal background*.

CONTEXT

WHAT DO WE MEAN BY FAIRNESS?

Fairness has different meanings to us depending on our *personal background*.

CONTEXT

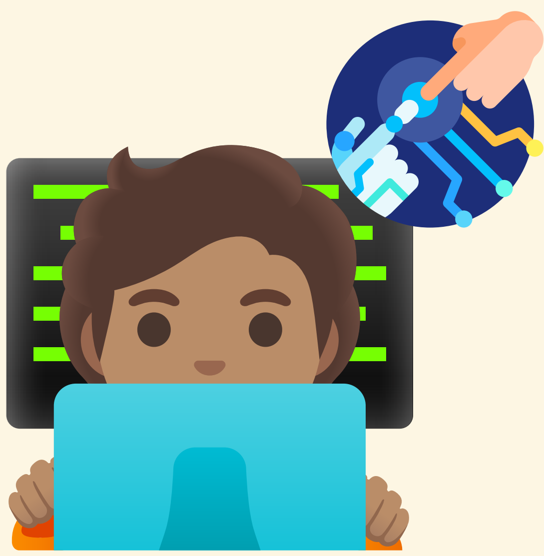
WHAT DO WE MEAN BY FAIRNESS?

Fairness has different meanings to us depending on our *personal background*.

CONTEXT

WHAT DO WE MEAN BY FAIRNESS?

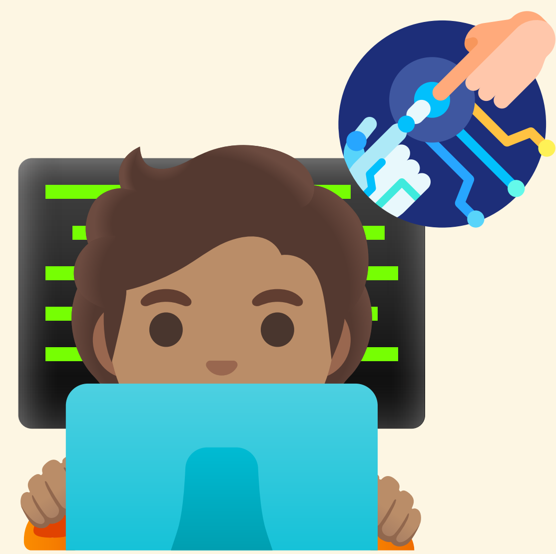
Fairness has different meanings to us depending on our *personal background*.



CONTEXT

WHAT DO WE MEAN BY FAIRNESS?

Fairness has different meanings to us depending on our *personal background*.



For people with predominantly scientific studies, fairness is something that should be **objectively measurable**. This is usually translated into the *fulfillment* of one or multiple fairness **metrics**.

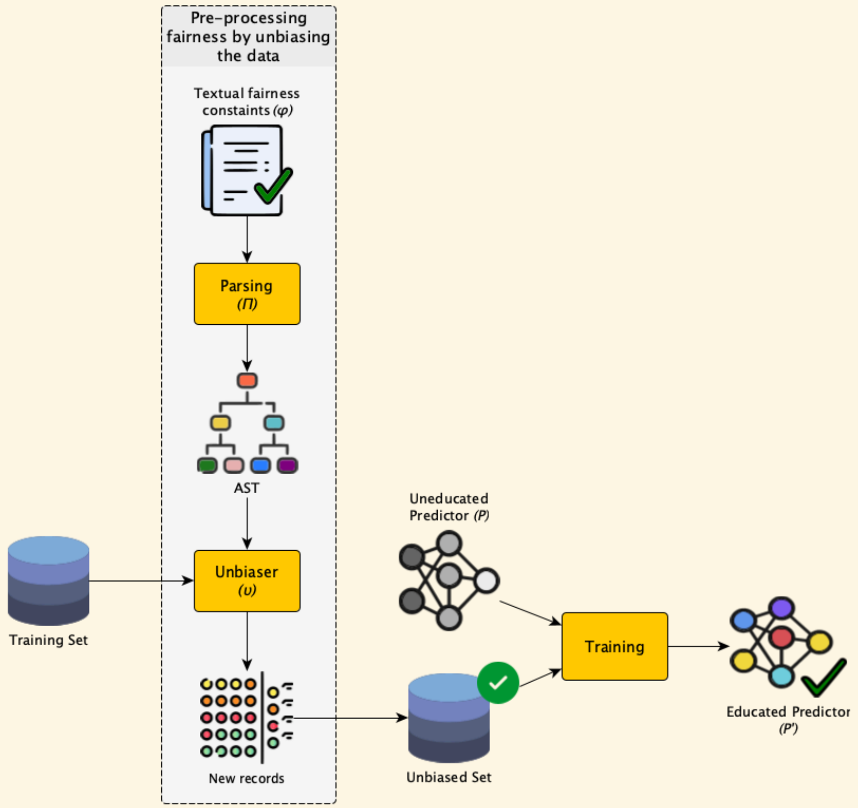
CONTEXT

ENFORCING FAIRNESS IN ML MODELS

CONTEXT

ENFORCING FAIRNESS IN ML MODELS

PRE-PROCESSING

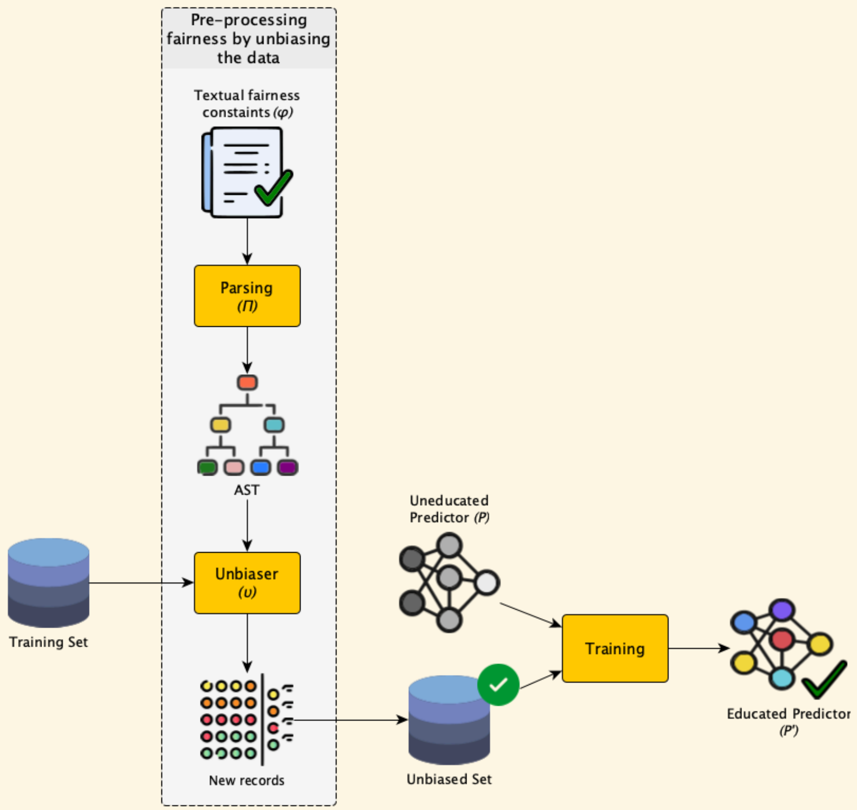


Methods that operate at **dataset level** to remove biases for sensitive groups.

CONTEXT

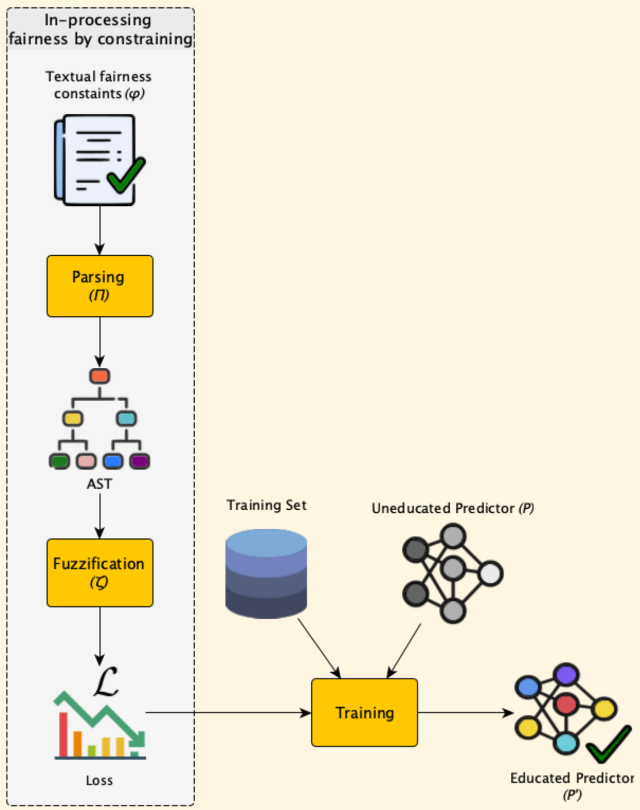
ENFORCING FAIRNESS IN ML MODELS

PRE-PROCESSING



Methods that operate at **dataset level** to remove biases for sensitive groups.

IN-PROCESSING

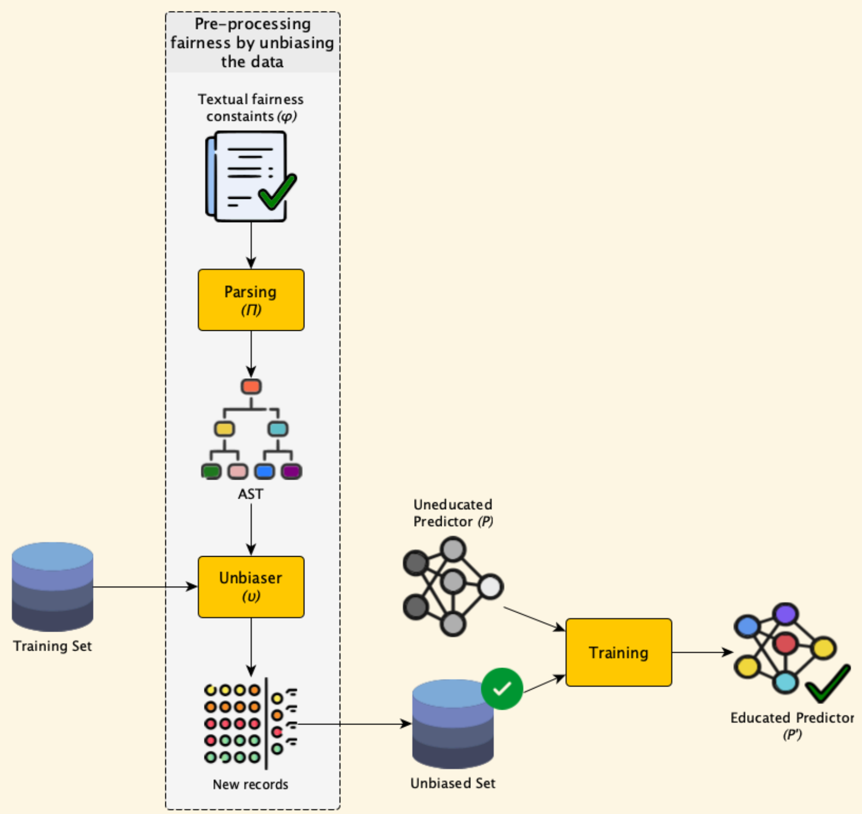


The **training of the model** takes into account the fairness constraints.

CONTEXT

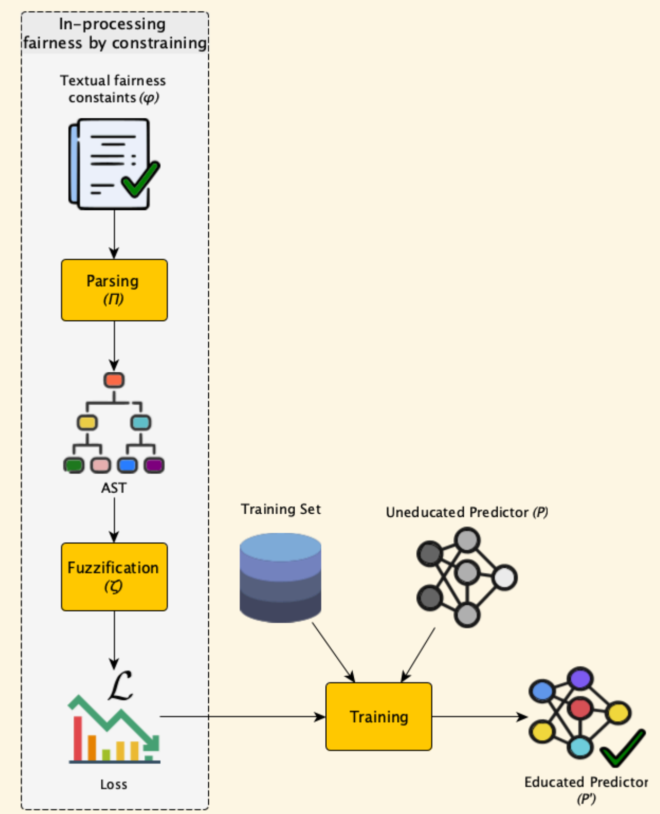
ENFORCING FAIRNESS IN ML MODELS

PRE-PROCESSING



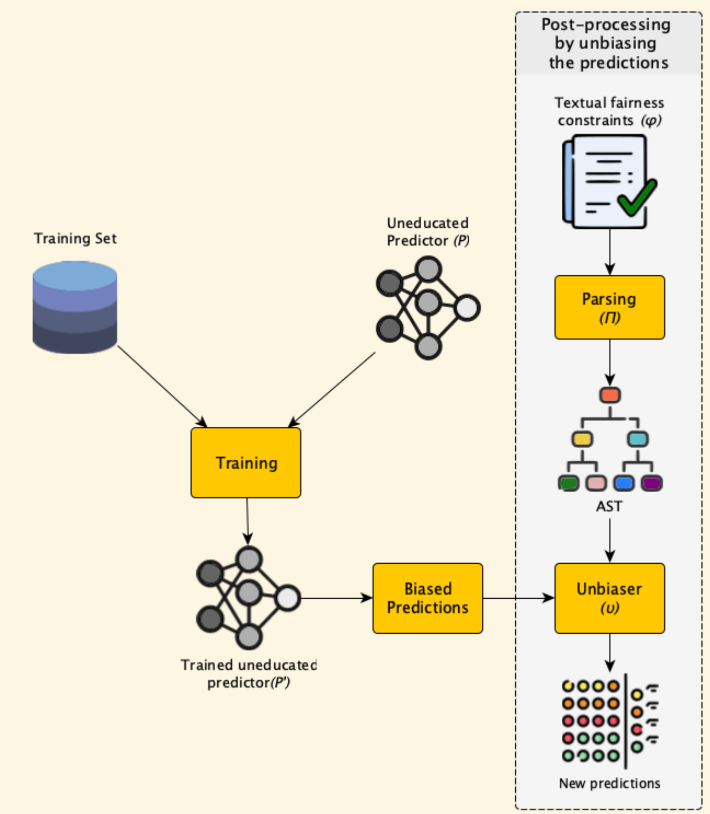
Methods that operate at **dataset level** to remove biases for sensitive groups.

IN-PROCESSING



The **training of the model** takes into account the fairness constraints.

POST-PROCESSING



The model is treated as a black-box and only the **predictions are adjusted** to ensure fairness.

CONTEXT

THE IN-PROCESSING TECHNIQUES

CONTEXT

THE IN-PROCESSING TECHNIQUES

PENALTY FUNCTION

A function, usually derived from a fairness metric, is chosen to *measure a violation* of fairness/bias. This function takes into account the **input data** and the **model's predictions**.

CONTEXT

THE IN-PROCESSING TECHNIQUES

PENALTY FUNCTION

A function, usually derived from a fairness metric, is chosen to *measure a violation* of fairness/bias. This function takes into account the **input data** and the **model's predictions**.

FUNCTION COMPUTATION

Because fairness metrics require **statistical distributions** to be computed, these distributions are *estimated on a subset (batch) of the data*. The actual computation of the fairness metric is therefore done during the loss computation.

CONTEXT

THE IN-PROCESSING TECHNIQUES

PENALTY FUNCTION

A function, usually derived from a fairness metric, is chosen to *measure a violation* of fairness/bias. This function takes into account the **input data** and the **model's predictions**.

FUNCTION COMPUTATION

Because fairness metrics require **statistical distributions** to be computed, these distributions are *estimated on a subset (batch) of the data*. The actual computation of the fairness metric is therefore done during the loss computation.

TRAINING

The loss function is a *combination* of the model's loss (e.g., binary cross entropy) and the fairness penalty. it is common to use a hyperparameter to balance the two terms.

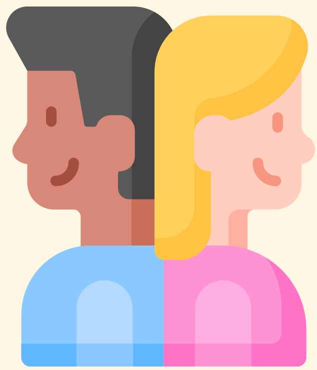
OPEN CHALLENGES

TYPES OF PROTECTED ATTRIBUTES

OPEN CHALLENGES

TYPES OF PROTECTED ATTRIBUTES

BINARY

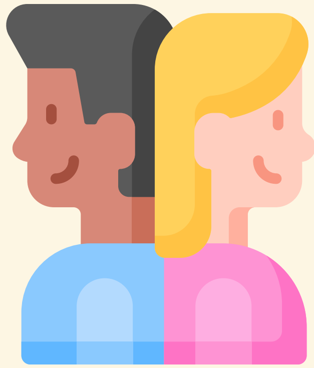


It is the **simplest case**, where the protected attribute can take only two values. There are only two groups to be considered, the classic example is the gender.

OPEN CHALLENGES

TYPES OF PROTECTED ATTRIBUTES

BINARY



It is the **simplest case**, where the protected attribute can take only two values. There are only two groups to be considered, the classic example is the gender.

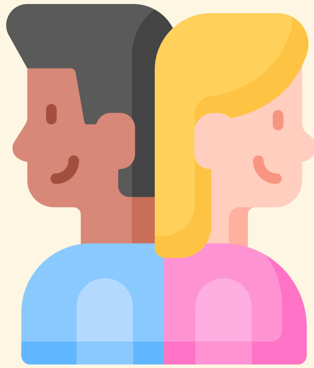
CATEGORICAL

The protected attribute can take more than two values. Here things start to get tricky, as we might **consider all the groups for fairness**. Examples are ethnicity, education, and occupation.

OPEN CHALLENGES

TYPES OF PROTECTED ATTRIBUTES

BINARY



It is the **simplest case**, where the protected attribute can take only two values. There are only two groups to be considered, the classic example is the gender.

CATEGORICAL

The protected attribute can take more than two values. Here things start to get tricky, as we might **consider all the groups for fairness**. Examples are ethnicity, education, and occupation.

CONTINUOUS

The protected attribute is a continuous variable. This is the *most complex case*, as we need to **estimate probability densities** to compute fairness metrics. An example is the income.

OPEN CHALLENGES

FAIRNESS METRICS

OPEN CHALLENGES

FAIRNESS METRICS

GROUP VS. INDIVIDUAL FAIRNESS

Group fairness is about **treating groups equally**, while individual fairness is about **treating similar individuals equally**.

Individual fairness metrics are more *computationally expensive* and because of that less common in practice.

However, also group fairness metrics can be computationally expensive. For this reason, we decided to focus on group fairness metrics.

OPEN CHALLENGES

FAIRNESS METRICS

GROUP VS. INDIVIDUAL FAIRNESS

Group fairness is about **treating groups equally**, while individual fairness is about **treating similar individuals equally**.

Individual fairness metrics are more *computationally expensive* and because of that less common in practice.

However, also group fairness metrics can be computationally expensive. For this reason, we decided to focus on group fairness metrics.

- *Demographic/statistical parity* how much model's predictions are **independent** of the protected attribute. $DP_{h,A}(X) = \sum_{a \in A} ||E[h(X) \mid A{=}a] - E[h(X)]||$
- *Disparate impact* how much the model **disproportionately** affects a group.

$$DI_{h,A}(X) = \min\left(\frac{E[h(X) \mid A{=}1]}{E[h(X) \mid A{=}0]}, \frac{E[h(X) \mid A{=}0]}{E[h(X) \mid A{=}1]}\right)$$

- *Equalized odds* how much the model **equally predicts** a given output for all the groups.

$$EO_{h,A}(X) = \sum_{(a,y)}^{A \times Y} eo_{h,A}(X, a, y)$$

$$eo_{h,A}(X, a, y) = ||E[h(X) \mid A{=}a, Y{=}y] - E[h(X) \mid Y{=}y]||$$

FAUCI

FAIRNESS UNDER CONSTRAINTS INJECTION

FAUCI

FAIRNESS UNDER CONSTRAINTS INJECTION

We design FaUCI in order to be *agnostic* to the fairness metric used and to the protected attribute type:

- we considered **demographic parity**, **disparate impact**, and **equalized odds** (any other metric can be used)
- we generalized the metric to work with **binary**, **categorical**, and **continuous** protected attributes
- we also considered ad-hoc weights for the groups to cover *corner cases* (e.g., strong imbalance)

FAUCI

FAIRNESS UNDER CONSTRAINTS INJECTION

We design FaUCI in order to be *agnostic* to the fairness metric used and to the protected attribute type:

- we considered **demographic parity**, **disparate impact**, and **equalized odds** (any other metric can be used)
- we generalized the metric to work with **binary**, **categorical**, and **continuous** protected attributes
- we also considered ad-hoc weights for the groups to cover *corner cases* (e.g., strong imbalance)

LOSS FUNCTION

$$L_{h,A}(X, Y) = E(h(X), Y) + \lambda F_{h,A}(X)$$

FAUCI

FAIRNESS UNDER CONSTRAINTS INJECTION

We design FaUCI in order to be *agnostic* to the fairness metric used and to the protected attribute type:

- we considered **demographic parity**, **disparate impact**, and **equalized odds** (any other metric can be used)
- we generalized the metric to work with **binary**, **categorical**, and **continuous** protected attributes
- we also considered ad-hoc weights for the groups to cover *corner cases* (e.g., strong imbalance)

LOSS FUNCTION

$$L_{h,A}(X,Y) = E(h(X),Y) + \lambda F_{h,A}(X)$$

BINARY AND CATEGORICAL

$$WDP_{h,A}(X) = \sum_{a \in A} ||E[h(X) \mid A{=}a] - E[h(X)]|| \cdot w_a$$

$$WDI_{h,A}(X) = \sum_{a \in A} \eta \left(\frac{E\left[h(X) \mid A{=}a\right]}{E\left[h(X) \mid A{\neq}a\right]} \right) \cdot w_a$$

$$WEO_{h,A}(X) = \sum_{(a,y)}^{A \times Y} eo_{h,A}(X,a,y) \cdot w_a$$

FAUCI

FAIRNESS UNDER CONSTRAINTS INJECTION

We design FaUCI in order to be *agnostic* to the fairness metric used and to the protected attribute type:

- we considered **demographic parity**, **disparate impact**, and **equalized odds** (any other metric can be used)
- we generalized the metric to work with **binary**, **categorical**, and **continuous** protected attributes
- we also considered ad-hoc weights for the groups to cover *corner cases* (e.g., strong imbalance)

LOSS FUNCTION

$$L_{h,A}(X,Y) = E(h(X),Y) + \lambda F_{h,A}(X)$$

BINARY AND CATEGORICAL

$$WDP_{h,A}(X) = \sum_{a \in A} ||E[h(X) \mid A{=}a] - E[h(X)]|| \cdot w_a$$

$$WDI_{h,A}(X) = \sum_{a \in A} \eta \left(\frac{E\left[h(X) \mid A{=}a\right]}{E\left[h(X) \mid A{\neq}a\right]} \right) \cdot w_a$$

$$WEO_{h,A}(X) = \sum_{(a,y)}^{A \times Y} eo_{h,A}(X,a,y) \cdot w_a$$

CONTINUOUS

$$GDP_{h,A}(X) = \int_l^u (||E[h(X) \mid A{=}a] - E[h(X)]|| \cdot w_a) \cdot da$$

$$GDI_{h,A}(X) = \int_l^u \eta \left(\frac{E\left[h(X) \mid A{=}a\right]}{E\left[h(X) \mid A{\neq}a\right]} \right) \cdot w_a \cdot da$$

$$GEO_{h,A}(X) = \int_l^u \sum_{(a,y)}^{A \times Y} (eo_{h,A}(X,a,0) + eo_{h,A}(X,a,1)) \cdot w_a \cdot da$$

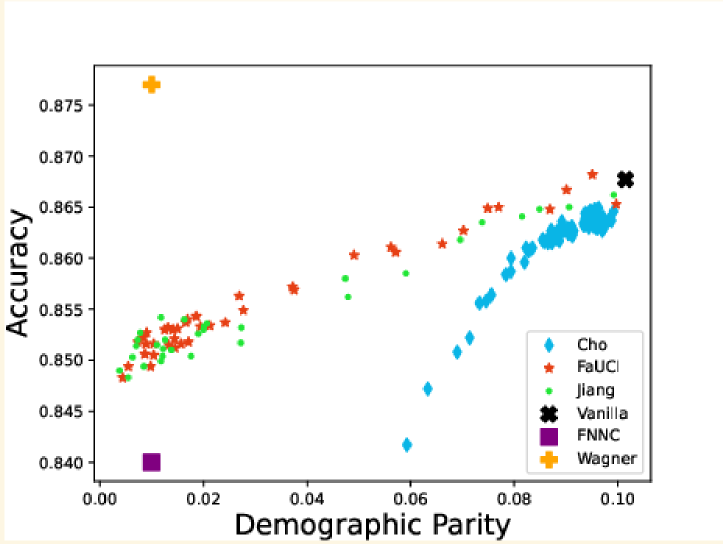
FAUCI

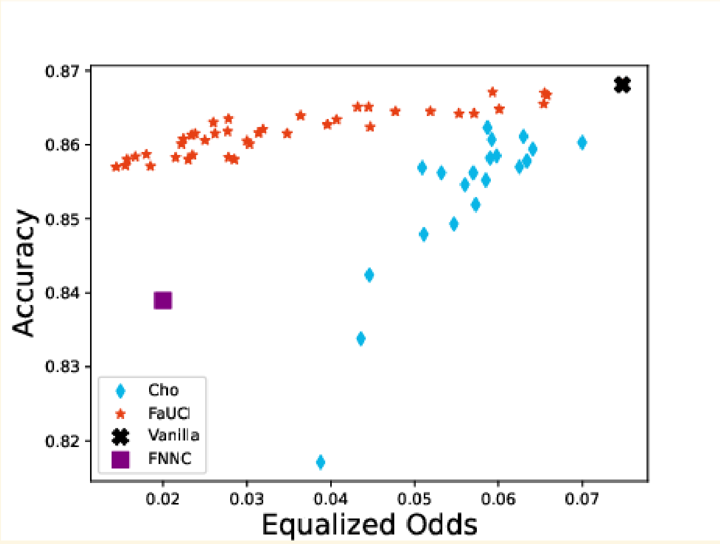
RESULTS ON THE ADULT DATASET

FAUCI

RESULTS ON THE ADULT DATASET

GENDER (BINARY)

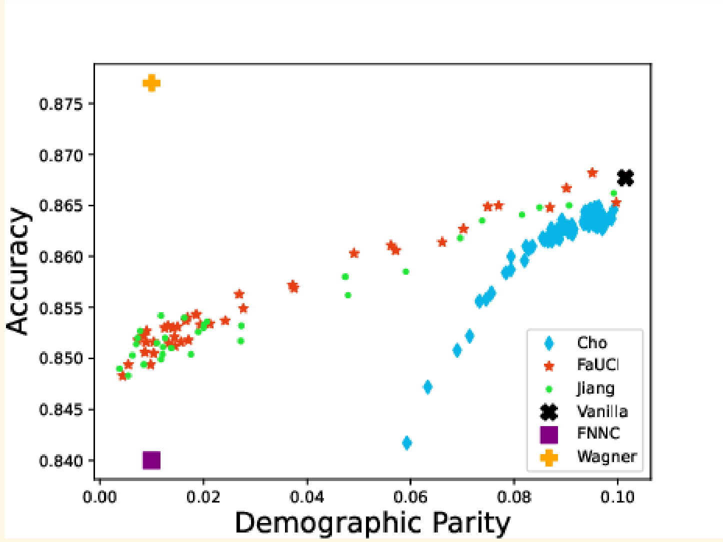


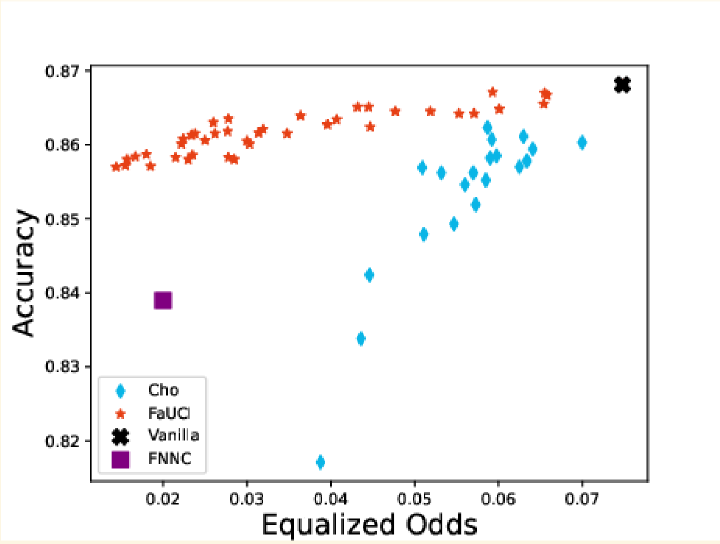


FAUCI

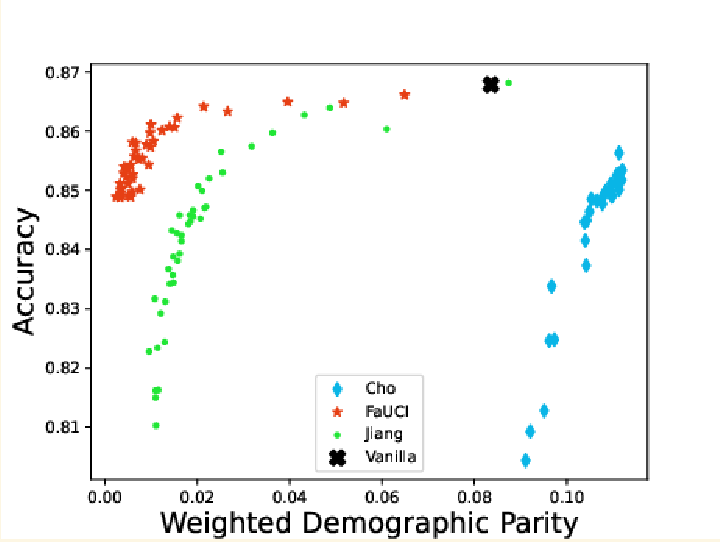
RESULTS ON THE ADULT DATASET

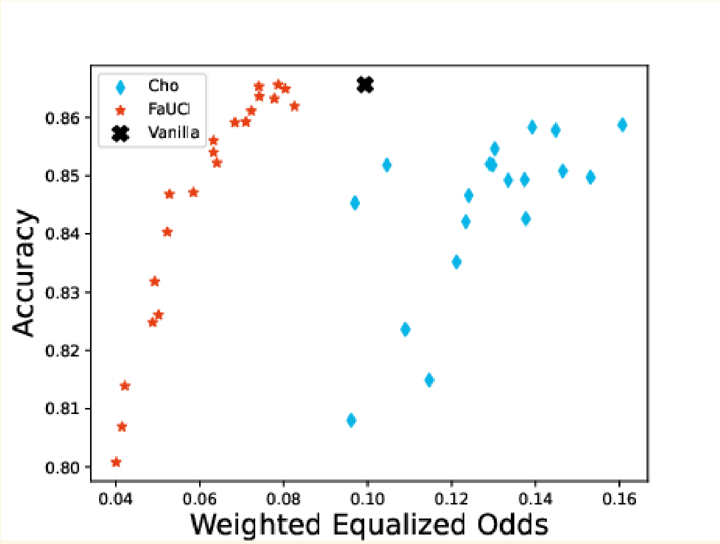
GENDER (BINARY)





ETHNICITY (CATEGORICAL)

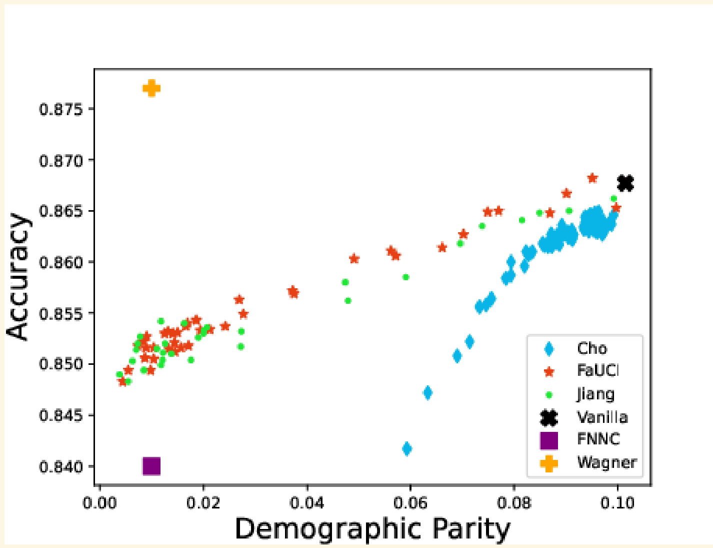


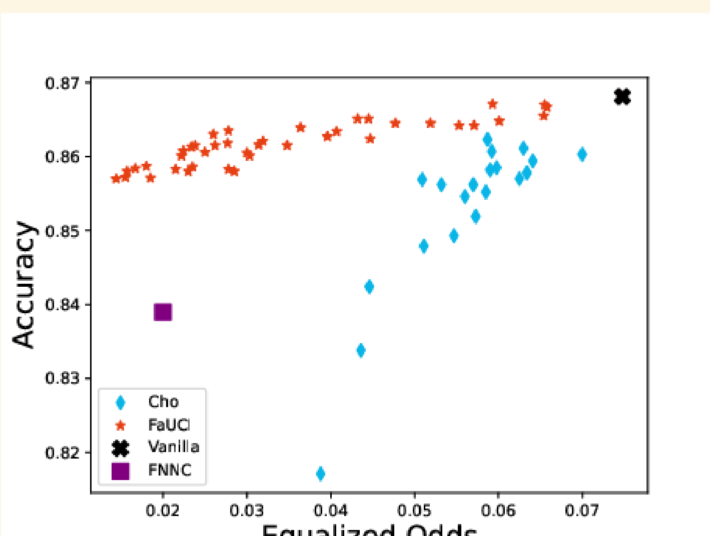


FAUCI

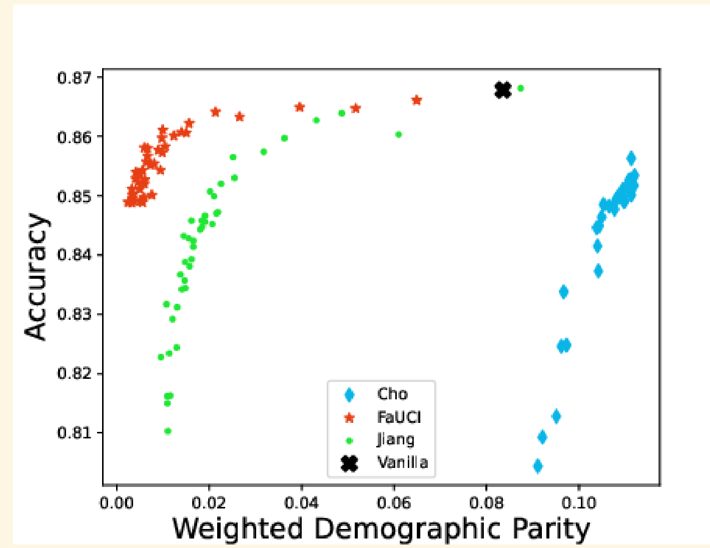
RESULTS ON THE ADULT DATASET

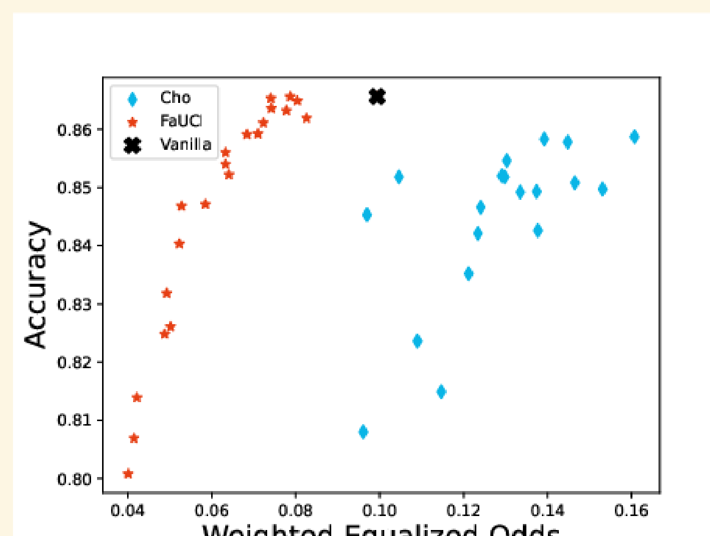
GENDER (BINARY)



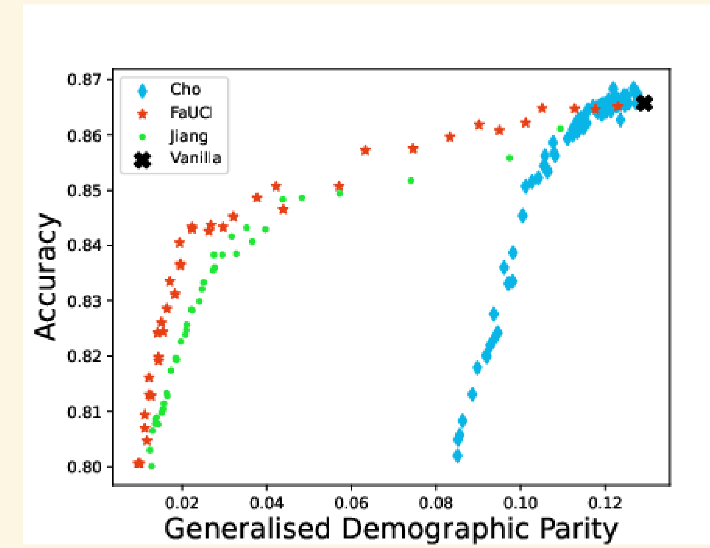


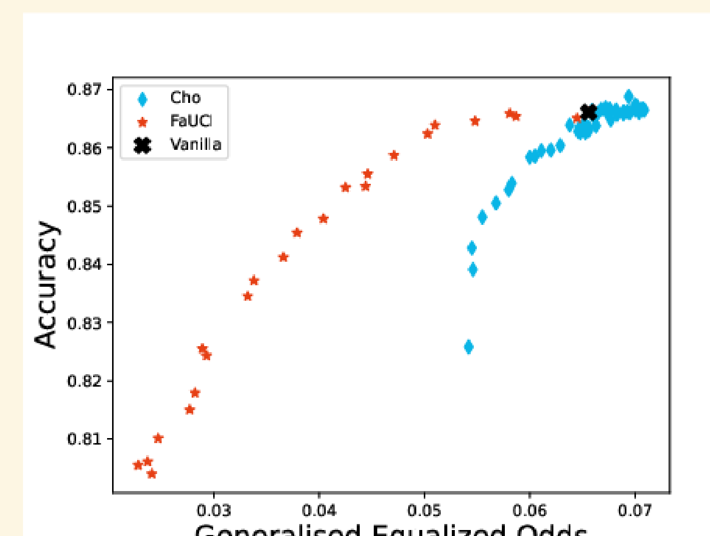
ETHNICITY (CATEGORICAL)





AGE (CONTINUOUS)





FUTURE DIRECTIONS

FUTURE DIRECTIONS

INTERSECTIONALITY

FaUCI can already be used to **consider multiple protected attributes** (subgroups) at the same time. However, we still need to perform a wide empirical study of the method to understand its performance.

$$L_{h,\bar{A}}(X,Y) = E(h(X),Y) + \lambda_1 F_{h,A_1}(X) + \cdots + \lambda_n F_{h,A_n}(X)$$

FUTURE DIRECTIONS

INTERSECTIONALITY

FaUCI can already be used to **consider multiple protected attributes** (subgroups) at the same time. However, we still need to perform a wide empirical study of the method to understand its performance.

$$L_{h,\bar{A}}(X,Y) = E(h(X),Y) + \lambda_1 F_{h,A_1}(X) + \cdots + \lambda_n F_{h,A_n}(X)$$

LANGUAGE FOR FAIRNESS

We want to develop a **language** to help users to define **ad-hoc fairness constraints** in a more intuitive way. Many potential users do not have a strong background in ML and statistics, so we aim to **make fairness techniques more accessible**. This is something very similar to what happen with *symbolic knowledge injection* methods.

FUTURE DIRECTIONS

INTERSECTIONALITY

FaUCI can already be used to **consider multiple protected attributes** (subgroups) at the same time. However, we still need to perform a wide empirical study of the method to understand its performance.

$$L_{h,\bar{A}}(X,Y) = E(h(X),Y) + \lambda_1 F_{h,A_1}(X) + \cdots + \lambda_n F_{h,A_n}(X)$$

LANGUAGE FOR FAIRNESS

We want to develop a **language** to help users to define **ad-hoc fairness constraints** in a more intuitive way. Many potential users do not have a strong background in ML and statistics, so we aim to **make fairness techniques more accessible**. This is something very similar to what happen with *symbolic knowledge injection* methods.

AUTOML FOR FAIRNESS

Because the training of ML models requires many hyperparameters – and with the addition of fairness constraints there is usually one more – we want to use AutoML tools to study the **convergence of the best hyperparameters** and how well they perform. In this way we can fairly compare different fairness techniques and understand which one is the best.

THANK YOU FOR YOUR ATTENTION!

AEQUITAS
unbias **AI**