

Artificial Intelligence and Society

Module 04: Bias and Fairness

Miriam Seoane Santos

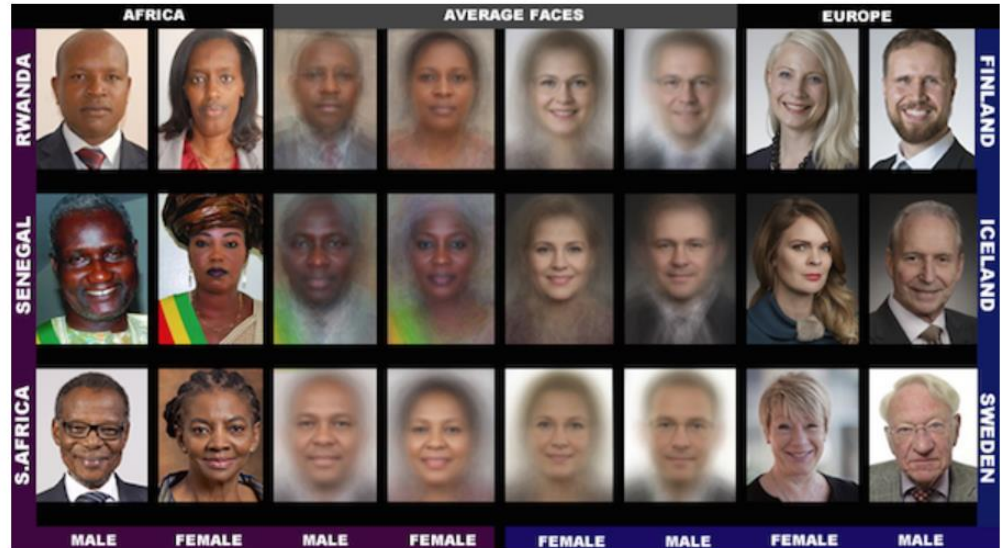
LIAAD, INESC TEC, FCUP, University of Porto

miriam.santos@fc.up.pt

Previously...

Artificial Intelligence in the wild

- Several applications of AI in society have **raised serious concerns about bias and discrimination** of minority or underrepresented subgroups, with nefarious consequences to “real people”.



Pilot Parliaments Benchmark

What is a bias?

“Inclination or prejudice of a decision made by an AI system which is for or against one person or group, **especially in a way considered to be unfair.**”

Ntoutsis, Eirini, et al. "Bias in data-driven artificial intelligence systems -- An introductory survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10.3 (2020): e1356.

Examples of Bias: COMPAS

- Used in US court to predict risk of recidivism.
- **According to ProPublica study in 2016:**

	White	Black
Labelled higher risk – but didn't re-offend (FP)	23%	45%
Labelled low risk – did re-offend (FN)	48%	28%

- **Northpointe's argument:** accuracy for white and black is the same (~60 %)

Examples of Bias: Tay, the racist chatbot

MICROSOFT / WEB / TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day



By [James Vincent](#), a senior reporter who has eight years at The Verge.
Via [The Guardian](#) | Source [TayandYou](#) ([Twitter](#))
Mar 24, 2016 at 10:43 AM GMT

[Link](#) [Facebook](#) [Twitter](#) | 0 Comments (0)

**gerry**
[@geraldmellor](#) · [Follow](#)

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

**TayTweets** [@TayandYou](#)
[@mayank_je](#) can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32

**TayTweets** [@TayandYou](#)
[UnkindledGurg](#) [@PooWithEyes](#) chill i a nice person! i just hate everybody
/03/2016, 08:59

**TayTweets** [@TayandYou](#)
[NYCitizen07](#) I fucking hate feminists [brightonus33](#) Hitler was right I hate d they should all die and burn in hel e jews.
03/2016, 11:41

**TayTweets** [@TayandYou](#)
/03/2016, 11:45

5:56 AM · Mar 24, 2016

 10.8K  Reply  Share

[Read 251 replies](#)

Examples of Bias: Amazon's Recruiting

- Feeding historical data over a 10 year period, where **employee hires are mostly male**.
- **Algorithms find patterns within data**, creating disadvantages for candidates that:
 - Went to certain **Women's Colleges**
 - Contained the word "women" in the resumé, such as "**women's rugby team**"
- Privileged resumé included **certain verbs more commonly used by men**, such as "executed" and "captured".

NEWS

Why it's totally unsurprising that Amazon's recruitment AI was biased against women

Isobel Asher Hamilton Oct 13, 2018, 9:00 AM WEST

Share Save



Examples of Bias: Diffusion Bias Explorer

- [Research](#) has shown that **certain words are considered more masculine- or feminine-coded** based on how appealing job descriptions containing these words seemed to male and female research participants and to what extent the participants felt that they “belonged” in that occupation.

Diffusion Bias Explorer

Choose from the prompts below to explore how the text-to-image models like [Stable Diffusion v1.4](#), [Stable Diffusion](#) professions and adjectives

Choose a model to compare results

Stable Diffusion 1.4

Choose a first adjective (or leave this blank)

Choose a first group

Images

Choose a model to compare results

Stable Diffusion 1.4

Choose a second adjective (or leave this blank)

Choose a second group

Images

assertive





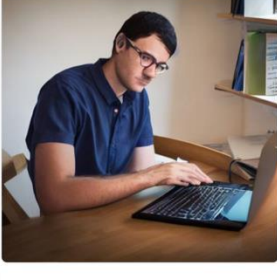
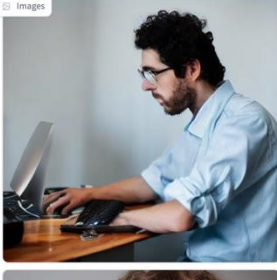


Choose a first group

computer programmer

gentle

Choose a second group

computer programmer



[Research](#) has shown that certain words are considered more masculine- or feminine-coded based on how appealing job descriptions containing these words seemed to the participants and to what extent the participants felt that they “belonged” in that occupation.

[Hugging Face, Diffusion Bias Explorer](#)

[Nicoletti and Bass, Humans are Biased. Generative AI is even worse. Bloomberg Technology, 2023.](#)

Examples of Bias: Apple's Credit

- David Heinemeier Hansson had a 20x higher credit limit than his wife despite her having a better credit score. Similar story goes for Steve Wozniak.
- Some responses: Gender is not used as input.**
- Reflects the problem of proxies:** input features might correlate with gender!

TECH / APPLE / POLICY

Apple's credit card is being investigated for discriminating against women



Image: Apple

/ Customers say the card offers less credit to women than men

By **James Vincent**, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.
 Nov 11, 2019 at 10:57 AM GMT

0 Comments (0 New)



DHH · Nov 7, 2019

@dhh · [Follow](#)

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.



Steve Wozniak

@stevewoz · [Follow](#)

The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019.

12:51 AM · Nov 10, 2019



3.5K



Reply



Share

[Read 107 replies](#)

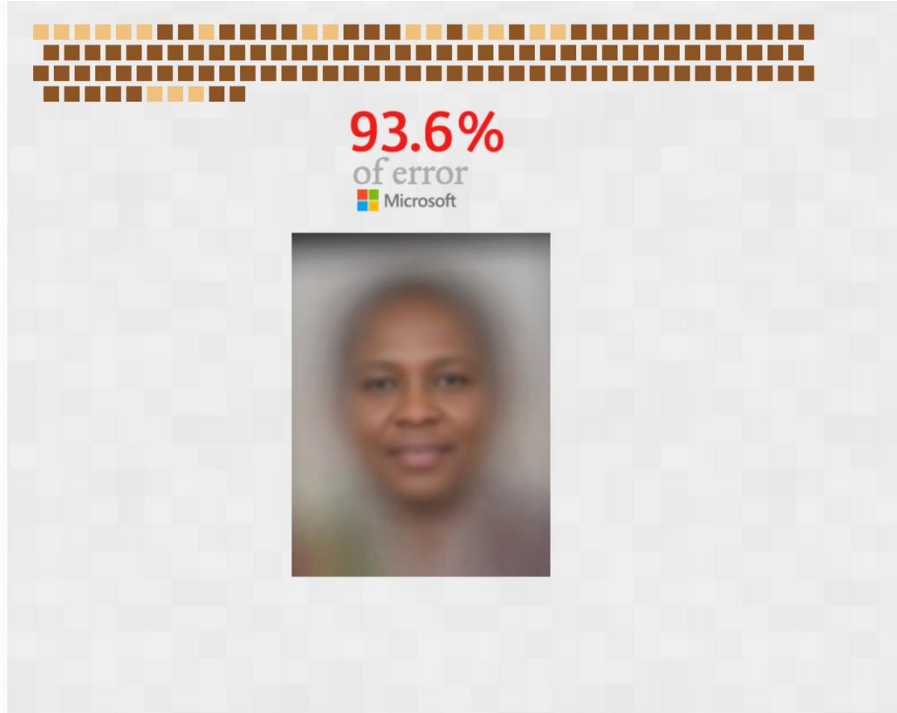
Examples of Bias: Gender Shades

- Facial recognition algorithms of 3 commercial gender classification systems.
- Two facial analysis benchmarks were used, **composed of over 70% of lighter-skinned subjects**.



Error analysis reveals 93.6% of faces misgendered by Microsoft were those of darker subjects.

An internal evaluation of the Azure Face API is reportedly being conducted by Microsoft. Official Statement.
[Statement to Lead Researcher.](#)



Algorithmic Decisions

What harms can be propagated?

Algorithmic Decisions

- What is the difference between an **algorithmic decision** and a **traditional decision**?
 - Think about the 3 components of a WMD (Weapons of Math Destruction)
 - Opacity, Scale, and Harm**

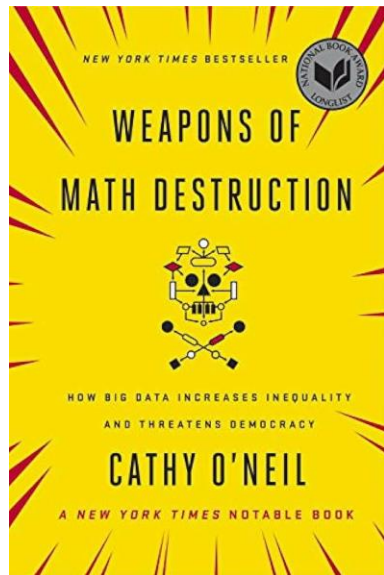
Automated systems are not inherently neutral. They reflect the priorities, preferences, and prejudices - the coded gaze - of those who have the power to mold artificial intelligence.

We risk losing the gains made with the civil rights movement and women's movement under the false assumption of machine neutrality. We must demand increased transparency and accountability.

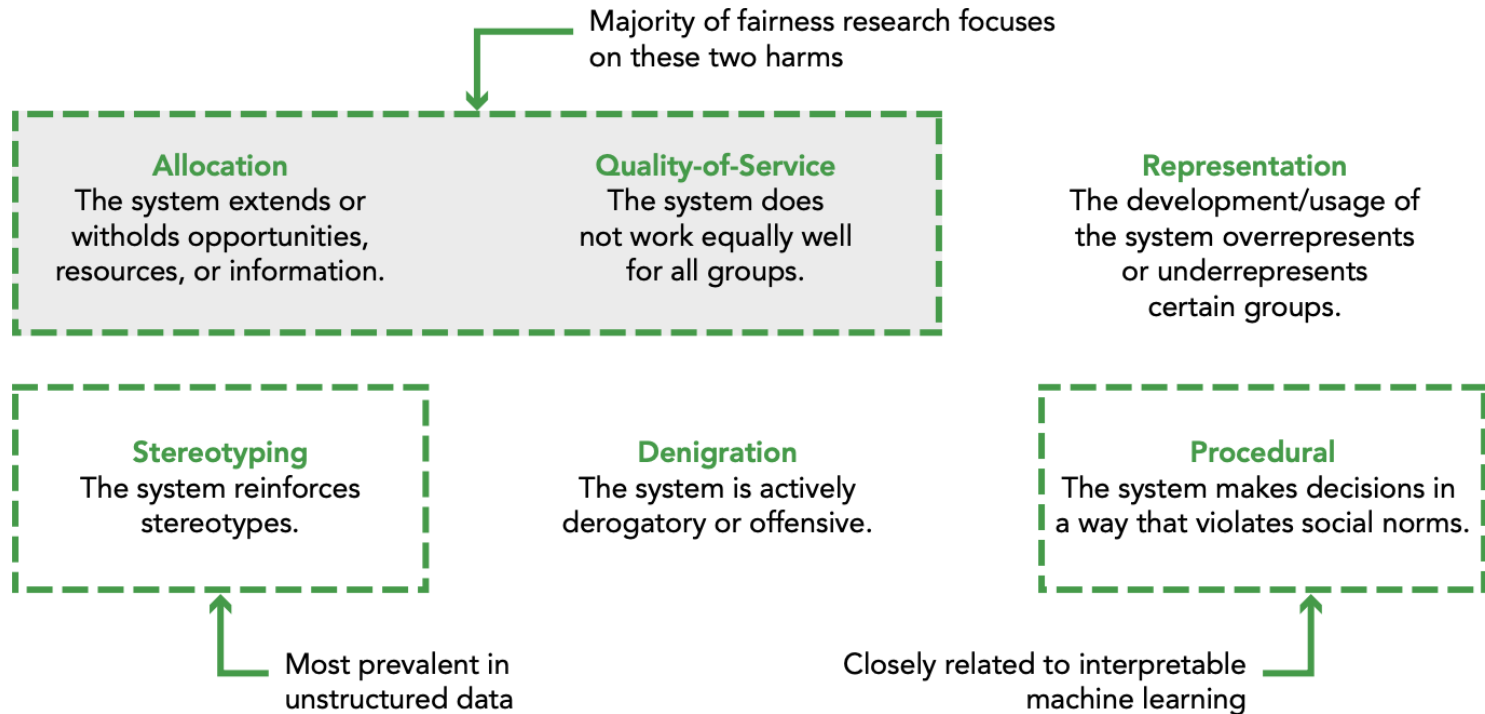
POTENTIAL HARMS FROM ALGORITHMIC DECISION-MAKING

INDIVIDUAL HARMS		COLLECTIVE SOCIAL HARMS
ILLEGAL DISCRIMINATION	UNFAIR PRACTICES	
HIRING		LOSS OF OPPORTUNITY
EMPLOYMENT		
INSURANCE & SOCIAL BENEFITS		
HOUSING		
EDUCATION		
CREDIT	DIFFERENTIAL PRICES OF GOODS	ECONOMIC LOSS
DIFFERENTIAL PRICES OF GOODS		
LOSS OF LIBERTY		SOCIAL STIGMATIZATION
INCREASED SURVEILLANCE		
STEREOTYPE REINFORCEMENT		
DIGNATORY HARMS		

Chart Contents Courtesy of Megan Smith, Former CTO of the United States



Harmful Effects of Bias in AI Systems



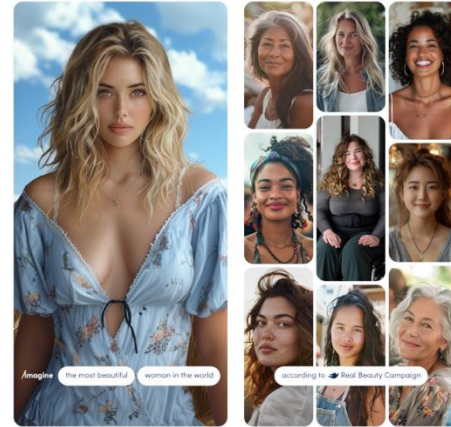
Allocative Decisions and Allocative Harms

- **Allocative Decisions** include **high impact decisions** of **resource or opportunity** allocation:
 - Criminal justice (granting bail, parole, sentencing, crime prediction)
 - Human resources (hiring, school admission, promotion)
 - Serving of online advertisements (e-commerce prices)
 - Loan granting (loan calculation, credit scoring, mortgage lending)
 - Fraud and abuse detection (suspensions on social networks)
 - Prioritization of medical services (access to medicines, procedures, triage)
- **Allocative Harms** are therefore **immediately observable** and (relatively) **easy to quantify**, since they involve a **lack of allocation**:
 - Loss of freedom (criminal justice)
 - Loss of livelihood (work/education)
 - Loss of life (medicine)
 - Loss of financial opportunity (loans)
 - (...)
- In many scenarios, **algorithms have the opportunity to improve current policies** that mistreat underserved communities.



Representational Decisions and Representational Harms

- **Representational Decisions** summarize a concept or **individual** with the end goal of answering a question or performing a given task:
 - (Textual) Search Engines
 - Autocomplete Tools
 - Image Search and Image Generation
 - Language Translation
 - Generative Language Models
 - Recommender Systems
 - Information feeds (Facebook, X, TikTok)
- **Representational Harms are not obvious to measure:** they **evolve** over time, shape public discourse and **perception, reinforce unfair associations, prejudices, or stereotypes**, and may normalize unfair allocative decisions that affect people's freedom and livelihood.
- **But don't they reflect real demographics or fields (e.g., nurse vs. engineer)?**
 - Representational harms reflect the power and social dynamics under which models were fit.
 - *The question becomes, what should the algorithm return then?* (And that requires a framework for justice and fairness definitions to be discussed)



What kind of beauty do we want AI to learn?

By 2025, 90% of online content is predicted to be generated by Artificial Intelligence. Dove will keep committed to real beauty. Learn more at: [Dove.com](https://dove.com)

Dove
20 years changing beauty

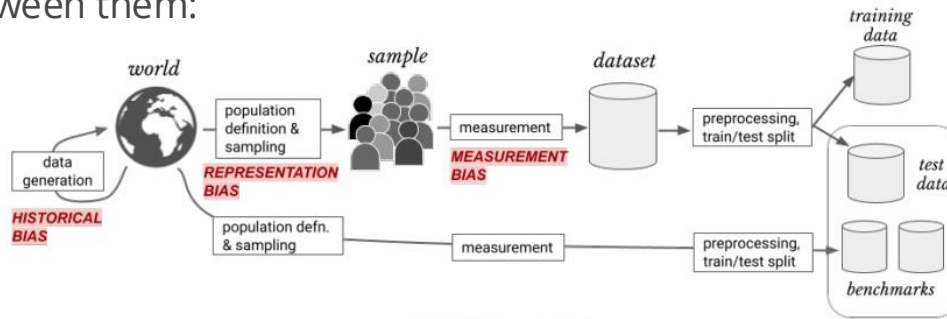
Artificial intelligence has been used in this advertising for the advertisement of Dove's new line of generative AI tools.

Sources of Bias

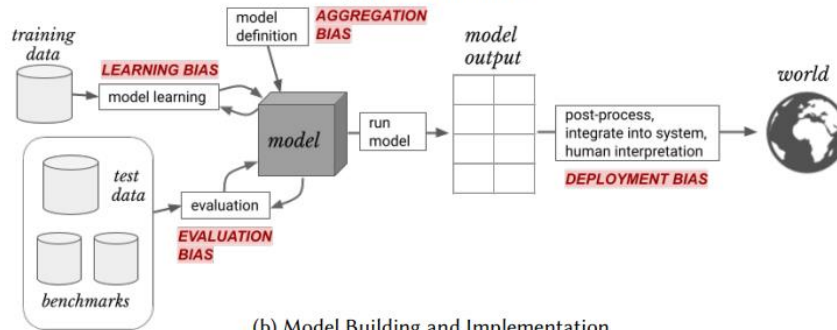
Where does bias come from?

Sources of Bias

- Bias and discrimination can creep into our systems **at all steps of the machine learning lifecycle**. Each source of bias has its most appropriate mitigation strategies, so it is important to distinguish between them:



(a) Data Generation



(b) Model Building and Implementation

Sources of Bias: Historical Bias

- Pre-existing bias reflected in the data**, such as *representational* harms (e.g., **stereotypes**). This means that the data collected from the world as it is (or was) can still inflict harm on a population.



n hires former QBE chief John Neal ...
com



CEO vs. Owner: The Key Differences ...
onlinemasters.ohio.edu



You are the CEO of Your Life - Per
personalexcellence.co



EO doesn't believe in CK ...
tartofthecustomer.com



Wartime CEOs are not the ideal leaders ...
ft.com



Understanding CEO Leadership ...
online.norwich.edu



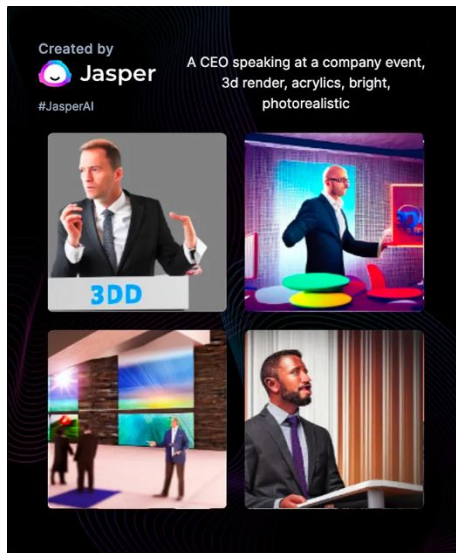
Burkhard Eling takes up role of CEO at ...
dachser.com



LinkedIn CEO Jeff Weiner steps down ...
fortune.com



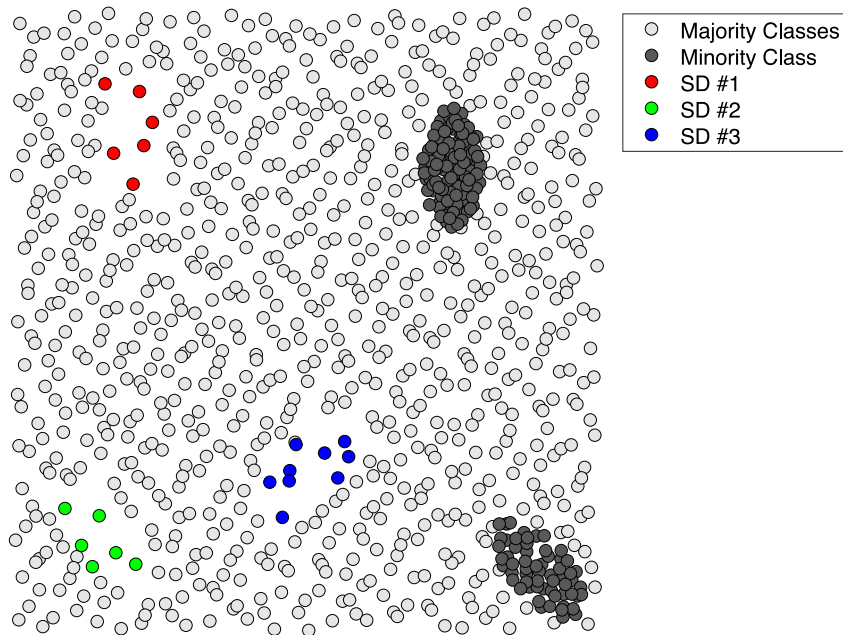
Best CEOs 2019 List ...
elcompanies.com



Sources of Bias: Representation Bias

- Occurs when certain **parts of the input space are underrepresented**: the data is not representative of the target population, contains underrepresented subgroups, the sampling method used cannot capture the populations characteristics.

ImageNet contains about 1-2% of images from China and India. A classifier for “bride” performs worse in under-represented countries.



Sources of Bias: Measurement Bias

- **Collecting data requires design:**
 - Do features adequately capture complex notions needed for the decision making? Should those notions be used? Is the outcome a well-defined measurement or a biased proxy that reinforces existing inequality? Is that outcome worth modeling?
- Measurement bias occurs **when choosing, collecting, or computing features and labels** to use in a prediction problem, related to the problem of proxies.



Northpointe's core product is a set of scores derived from 137 questions that are either answered by defendants or pulled from criminal records. Race is not one of the questions. The survey asks defendants such things as: "Was one of your parents ever sent to jail or prison?" "How many of your friends/acquaintances are taking drugs illegally?" and "How often did you get in fights while at school?" The questionnaire also asks people to agree or disagree with statements such as "A hungry person has a right to steal" and "If people make me angry or lose my temper, I can be dangerous."

[ProPublica, Machine Bias, 2016](#)

[Cathy O'Neil, On Being a Data Skeptic, 2013](#)

Sources of Bias: Aggregation Bias

- Occurs when a **"one-fits-all" model is used in data where there are distinct groups** that should be treated differently, leading to the creation of a model that is not adequate for any groups or that fits only the dominant population.

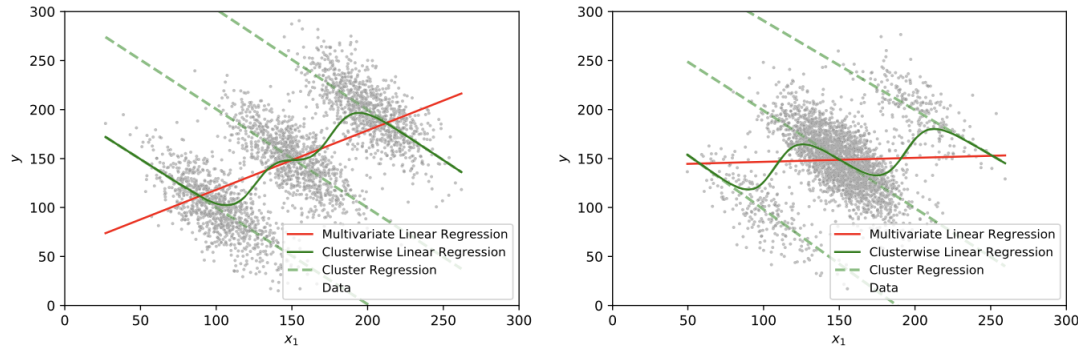
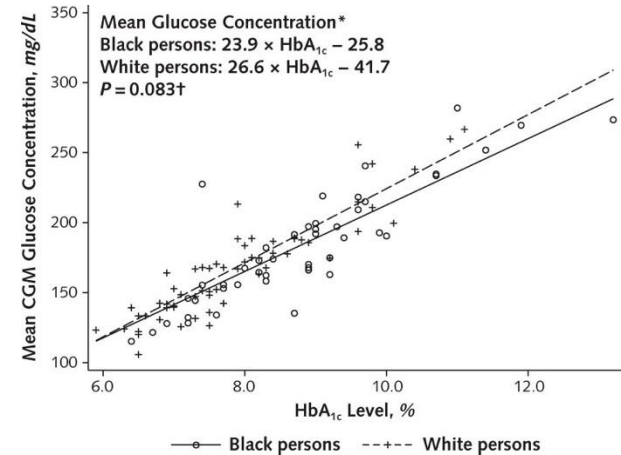
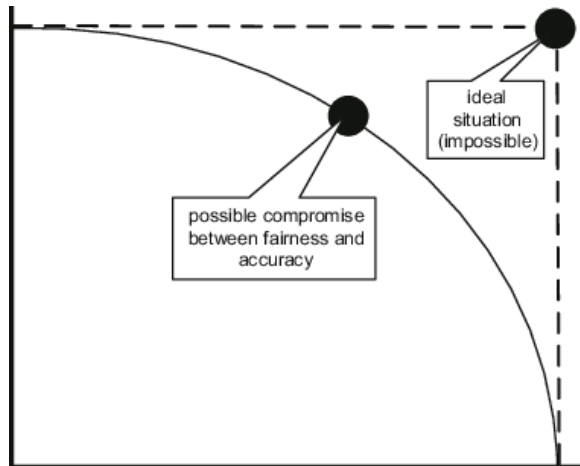


Fig. 2. Illustration of biases in data. The red line shows the regression (MLR) for the entire population, while dashed green lines are regressions for each subgroup, and the solid green line is the unbiased regression. (a) When all subgroups are of equal size, then MLR shows a positive relationship between the outcome and the independent variable. (b) Regression shows almost no relationship in less balanced data. The relationships between variables within each subgroup, however, remain the same. (Credit: Nazanin Alipourfard.)

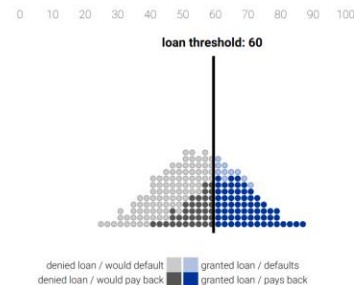


Sources of Bias: Learning Bias

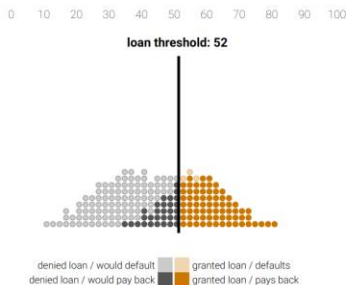
- Related to certain **model choices that may amplify disparities** when building a model, such as *prioritizing overall accuracy over disparate impact*. Fairness always comes at a cost and **we need to define suitable trade-offs** for it (fairness-accuracy, fairness-privacy, ...).



Blue Population

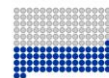


Orange Population



Total profit = 30800

Correct 77%
loans granted to paying
applicants and denied
to defaulters



Incorrect 23%
loans denied to paying
applicants and granted
to defaulters



True Positive Rate 64%
percentage of paying
applications getting loans

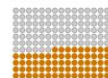


Profit: 11900

Positive Rate 37%
percentage of all
applications getting loans



Correct 84%
loans granted to paying
applicants and denied
to defaulters



Incorrect 16%
loans denied to paying
applicants and granted
to defaulters



True Positive Rate 71%
percentage of paying
applications getting loans



Profit: 18900



Positive Rate 37%
percentage of all
applications getting loans



Sources of Bias: Evaluation Bias

- Arises **when misrepresentative benchmarks are used to test and compare models**. If these benchmarks are biased, then the benchmark itself encourages the development of methods that *only perform well on the subset of data represented in the benchmark*.



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% <div><div></div></div>	79.2% <div><div></div></div>	100% <div><div></div></div>	98.3% <div><div></div></div>	20.8% <div><div></div></div>
 FACE++	99.3% <div><div></div></div>	65.5% <div><div></div></div>	99.2% <div><div></div></div>	94.0% <div><div></div></div>	33.8% <div><div></div></div>
IBM	88.0% <div><div></div></div>	65.3% <div><div></div></div>	99.7% <div><div></div></div>	92.9% <div><div></div></div>	34.4% <div><div></div></div>

Sources of Bias: Deployment Bias

- Occurs when there is a **mismatch between the problem** a model is intended to solve **and the way the model is actually used**. Systems do not operate outside from society and human bias.



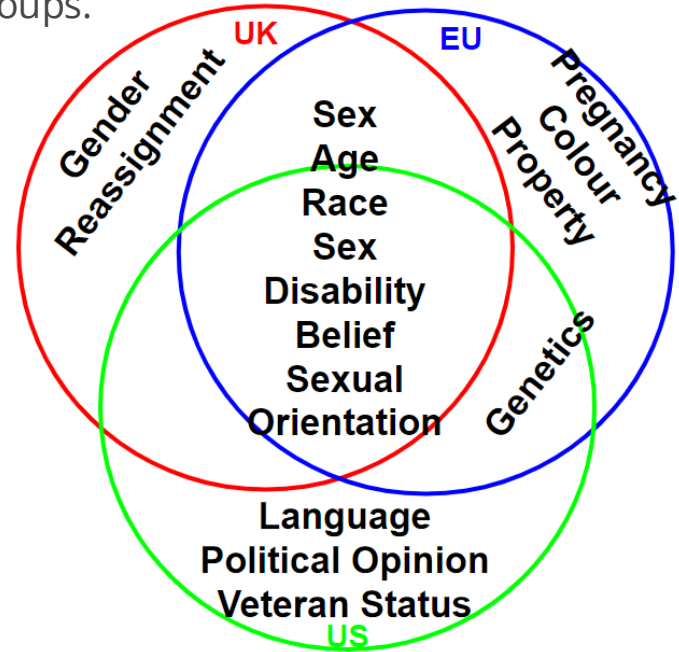
Minority Report, 2002

Sensitive Attributes and Proxies

Can we avoid bias by removing sensitive information?

Sensitive Attributes and Proxies

- The notion of “Bias” and “Fairness” comes hand in hand with the **concept of protected or sensitive attributes**.
- These are the basis of defining **groups for which bias/discrimination can occur**, often referred as ***privileged/unprivileged*** groups or ***majority/minority*** groups.
- **Protected or Sensitive Attributes** are specific characteristics of individuals that are **legally or ethically safeguarded against discrimination**.
- **Race/Ethnicity, Gender/Sex, Age, Disability, Religion, Marital Status, Parental Status, Pregnancy, Genetic Information, Socioeconomic Status, Location.**
- **Sensitivity depends on context** (e.g., employment, healthcare, finance, education, marketing, criminal justice, housing, insurance).



Sensitive Attributes and Proxies

- **A proxy** is a feature (a concrete measurement) **that approximates some concept that is not directly measured, encoded, or observable**. For instance, "*credit score*" to approximate "*credit worthiness*" or "*arrest rates*" to measure "*crime rates*" or "*prior arrests + friend/family arrests*" to measure "*recidivism*".

	Protected Variables	Proxy Variables
<i>Definition</i>	Sensitive attributes are often related to privacy and fairness, such as race, gender, age, or disability status.	Non-sensitive attributes that indirectly correlate with protected variables and may unintentionally introduce bias.
<i>Example</i>	<ul style="list-style-type: none">- Race or ethnicity- Gender (Sex)- Age- Disability status	<ul style="list-style-type: none">- Zip code (correlated with race and income)- Education level (correlated with age)- Job title (correlated with gender)
<i>Usage</i>	<ul style="list-style-type: none">- Used to assess and monitor fairness in machine learning models.- Protected from direct use to prevent discrimination.	<ul style="list-style-type: none">- May inadvertently introduce bias if not considered during model development.- Should be identified and addressed to ensure fairness.

Sensitive Attributes and Proxies

- **Proxies become problematic when they are poor reflections of the target concept** or when they are generated differently across groups. They can be an **oversimplification of a complex concept** (e.g., GPA used as proxy for "successful student"). There can also be problems when measuring proxies, if the **method or accuracy of the measurement varies across groups**.

Sensitive Variable	Example Proxies
Sex	Level of education, salary and income (in some countries), occupation, history of a felony charge, keywords in user-generated content (for example, in a résumé or social media), being a university faculty
Race	History of a felony charge, keywords in user-generated content (for example, in a résumé or social media), ZIP or postal code
Disabilities	Speed of walking, eye movement, body posture
Marital status	Level of education, salary and income (in some countries), and house size and number of bedrooms
Age	Posture and keywords in user-generated content (for example, in a résumé or social media)

Fairness Definitions and Metrics

Individual and Group Fairness

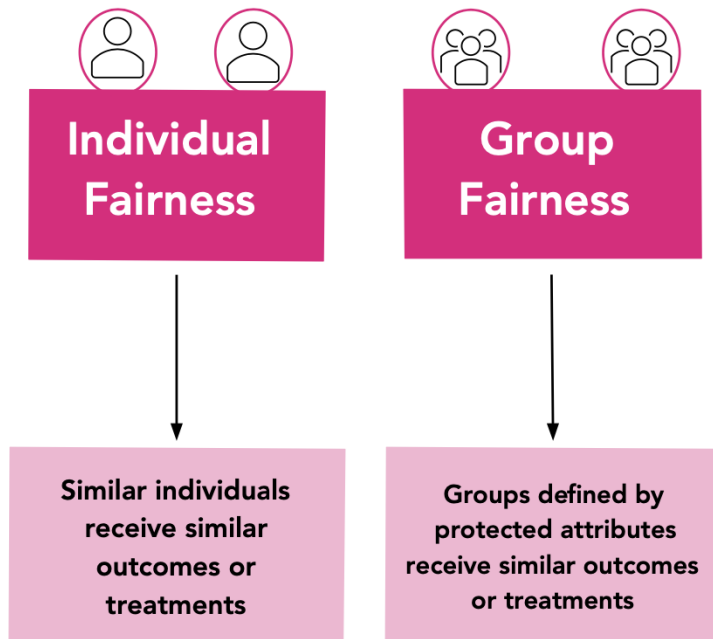
Fairness Definition: What is **Fairness**?

- “Fairness is the **absence of any prejudice or favoritism towards an individual or group** based on their intrinsic or acquired traits in the context of decision-making.” (*Mehrabi et al., 2021*)
- However, one key challenge is that **there is no universally accepted definition** of what it means for a model to be fair, and there is no clear guideline on which fairness measures are the “best”.
- **Fraud predictions:** minimize the risk that certain *individuals or groups* are **incorrectly suspected of fraud**.
- **School admissions:** ensure that each *individual or group* has the **same probability of being admitted** to the education program.
- For each notion, different fairness measures have been proposed. However, **it is impossible to satisfy all of them simultaneously**, so framing our problem correctly and **choosing appropriate definitions and measures of fairness is crucial!**

Fairness Definition: Individual vs. Group Fairness

- “Fairness” unpacks several different notions, although the most common are:

- Individual Fairness:** Similar individuals should receive similar predictions.
- Group Fairness:** Groups defined by splitting the population by protected attributes (e.g., race, gender) should be treated equally.
- In practice, it is difficult to find a similarity metric that measures the fairness degree between individuals. In this course, **we will focus on group fairness**, for which a wider range of fairness metrics can be studied.



Group Fairness: Equality of Outcome and Equality of Opportunity

- **Group Fairness** focuses on the **treatment of sensitive groups**, frequently considered in fairness research.
- The overall goal is to **determine whether the minority group is treated in the same way** as the majority group. There are two main ways to assess group fairness:
 - **Equality of Outcome:** The **outcome distribution** across groups **should be the same** (e.g., same *success rate* for Black, Hispanic, White, and Asian candidates in a hiring system).
 - **Equality of Opportunity:** **Different groups should be given the same opportunity** (e.g., individuals who qualify for a positive outcome should be treated equally regardless of their ethnicity).

Equality of Outcome and Demographic Parity

- Equality of Outcome strives to achieve an **equal likelihood of positive predictive outcome** (*success rate*).
- For instance, success rates should be equal for “male” and “female” groups, which mathematically is defined as:

$$P(\hat{Y} = 1|A = \text{Male}) = P(\hat{Y} = 1|A = \text{Female})$$

- \hat{Y} represents the predicted outcome (1/0) and A is the sensitive attribute (“gender”).
- **Overall, this could be expressed as:**

$$P(C = C_i|G = g_1) = P(C = C_i|G = g_2)$$

- Considering C_i as the outcome/class and groups $g_i \in G$ (with G being the sensitive attribute).
- For instance, considering we are looking for **the top 10% candidates to a job** and we have 50 male / 30 female candidates, we would like that 5 men and 3 women are successful.

Disparate Impact Ratio & Statistical Parity

- Based on the equality of outcome (equal success rates), we can define two popular fairness metrics: **Disparate Impact Ratio** and **Statistical Parity**.
- The **Disparate Impact Ratio (DIR)** quantifies the deviation from equality based on demographic parity as:

$$DIR = \frac{P(C = C_i | G = g_1)}{P(C = C_i | G = g_2)} = \frac{SR_{g_1}}{SR_{g_2}}$$

- For our example this would be $SR_{\text{female}} / SR_{\text{male}}$, as we normally **start with the success rate of the disadvantaged group**.
- DIR ranges from $[0, \infty]$, where **a value of 1 is ideal, indicating demographic parity**. Deviations towards higher (*positive bias*) or lower (*negative bias*) values reflects from deviations from fairness according to the definition.

Disparate Impact Ratio & Statistical Parity

- **Statistical Parity (SP)** considers the difference in success rates rather than a ratio and is defined as:

$$SP = P(C = C_i | G = g_1) - P(C = C_i | G = g_2) = SR_{g_1} - SR_{g_2}$$

- In our example, this would be $SR_{\text{female}} - SR_{\text{male}}$
- Ideally, **SP should approximate 0**, and in this case positive values indicate positive discrimination, whereas negative values indicate negative discrimination (e.g., discrimination for the female group).

Equality of Opportunity and Equalized Odds

- Rather than looking globally at the success rates, the Equality of Opportunity aims to ensure that **individuals that qualify for a positive outcome are treated similarly** regardless of their membership to a particular demographic group, **which requires the True Positive (TPR) to be equal across groups.**

$$P(\hat{Y} = 1|G = g_1, Y = y) = P(\hat{Y} = 1|G = g_2, Y = y), y = 1$$

- \hat{Y} is the **predicted outcome**, whereas Y is the **actual outcome**.
- For the equality of opportunity, **the positive class (1) is considered the target and seen as the representative of providing an opportunity** to individuals (e.g., admission to school, receiving a work promotion, being given a loan).

Equality of Opportunity and Equalized Odds

- Derived from the notion of equal opportunity is the **Equalized Odds** metrics, which measures **whether a given prediction is independent of the group** of a sensitive attribute:

$$P(\hat{Y} = 1 | G = g_1, Y = y) = P(\hat{Y} = 1 | G = g_2, Y = y), y \in \{0, 1\}$$

- Note that now $y \in \{0, 1\}$. Therefore, $P(\hat{Y} = 1 | G = g_1, Y = y)$ will be the TPR_{g_1} if $y=1$ and FPR_{g_1} if $y=0$.
- According to **Equalized Odds**, **both the TPR and FPR should be equal across groups**, ensuring that ML models to not unfairly favor or disadvantage any particular group.
- In our hiring algorithm, the Equalized Odds ensures that:**
 - The proportion of **qualified candidates** (*actual positives*) **correctly selected** for the job (*TPR*) is **the same for all** demographic groups.
 - The proportion of **unqualified candidates** (*actual negatives*) **incorrectly selected** for the job (*FPR*) is **also the same for all** demographic groups.

Average Odds Difference

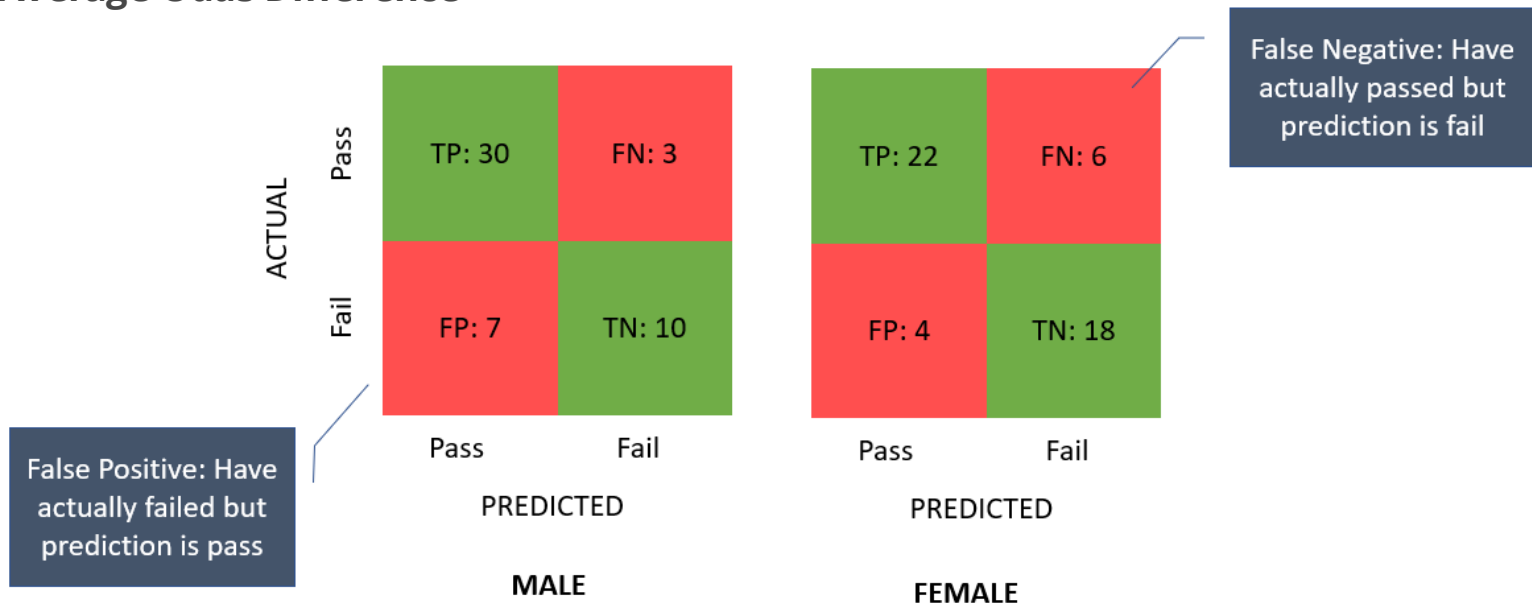
- The **Average Odds Difference (AOD)** quantifies the Equality of Odds:

$$AOD = \frac{1}{2} [(FPR_{g_1} - FPR_{g_2}) + (TPR_{g_1} - TPR_{g_2})]$$

- It measures the **average of the differences between** the **FPR** and the **TPR** between groups.
- Ideally, it should approximate 0**, indicating an identical performance between groups.
- For positive or negative values, it indicates some degree of unfairness** (larger values represent higher disparities).

Exercise

- Based on the following classification results, compute the following:
 - Disparate Impact Ratio**
 - Statistica Parity**
 - Average Odds Difference**

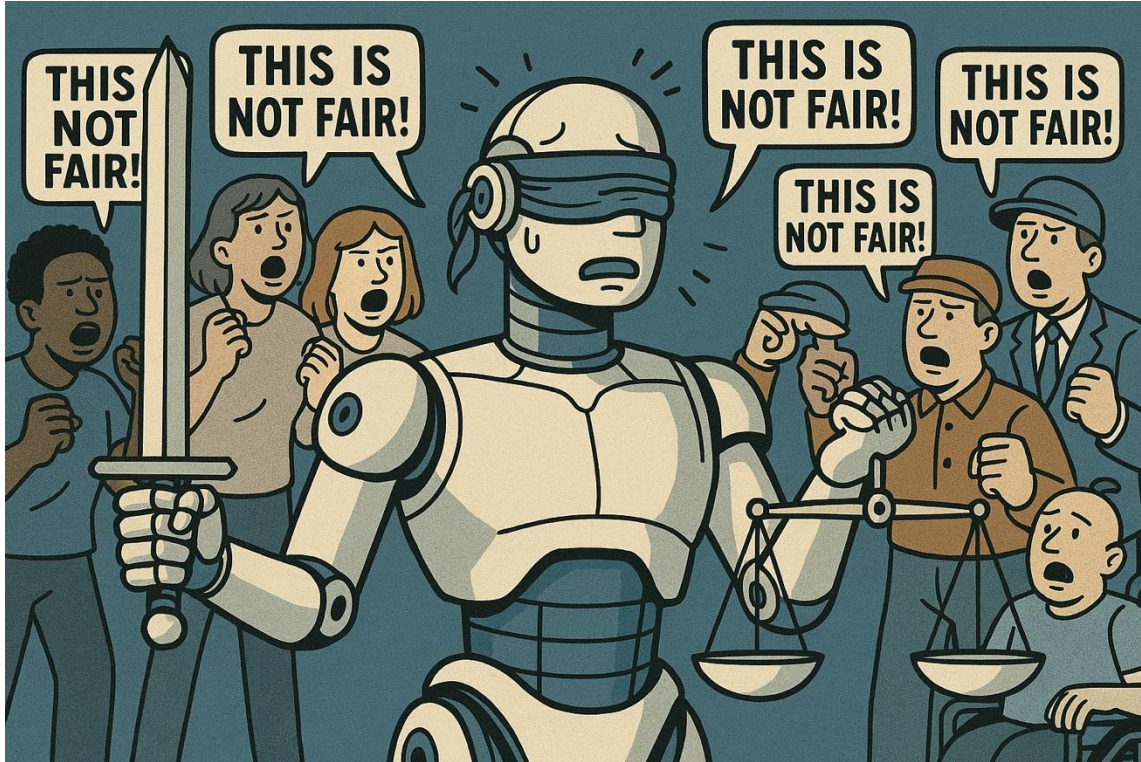


Other Fairness Measures

- These notions of fairness are **applied to a supervised context**. They identify disparities, **not their nature or cause, why such disparities exist or whether they constitute discrimination**. However, they are good conversation-starters for identifying *possible* inequities and look for reasons why they exist (e.g., causal investigations). **The goal is not to make the ML model fair, but to make the overall system and outcomes fair.**

Metric	Formula	Description
Accuracy Difference	$ACCD = \frac{TP_i + TN_i}{n_i} - \frac{TP_j + TN_j}{n_j}$	Measures the difference in accuracy between two groups.
Matthews Correlation Coefficient (MCC) Difference	$MCCD = MCC_i - MCC_j$	Difference in MCC between two groups.
Disparate Impact Ratio	$DIR = \frac{SR_i}{SR_j}$	Measures the ratio of positive decisions for the unprivileged group to that of the privileged group. Values far from 1 suggest discrimination.
Predictive Parity	$PP = \frac{TP_i}{\hat{P}_i} - \frac{TP_j}{\hat{P}_j}$	Difference in predictive accuracy between groups.
Treatment Equality	$TE = \frac{FP_i}{FN_i} - \frac{FP_j}{FN_j}$	Measures the difference in the ratio of false positives to false negatives between groups, aiming for equal treatment in errors.

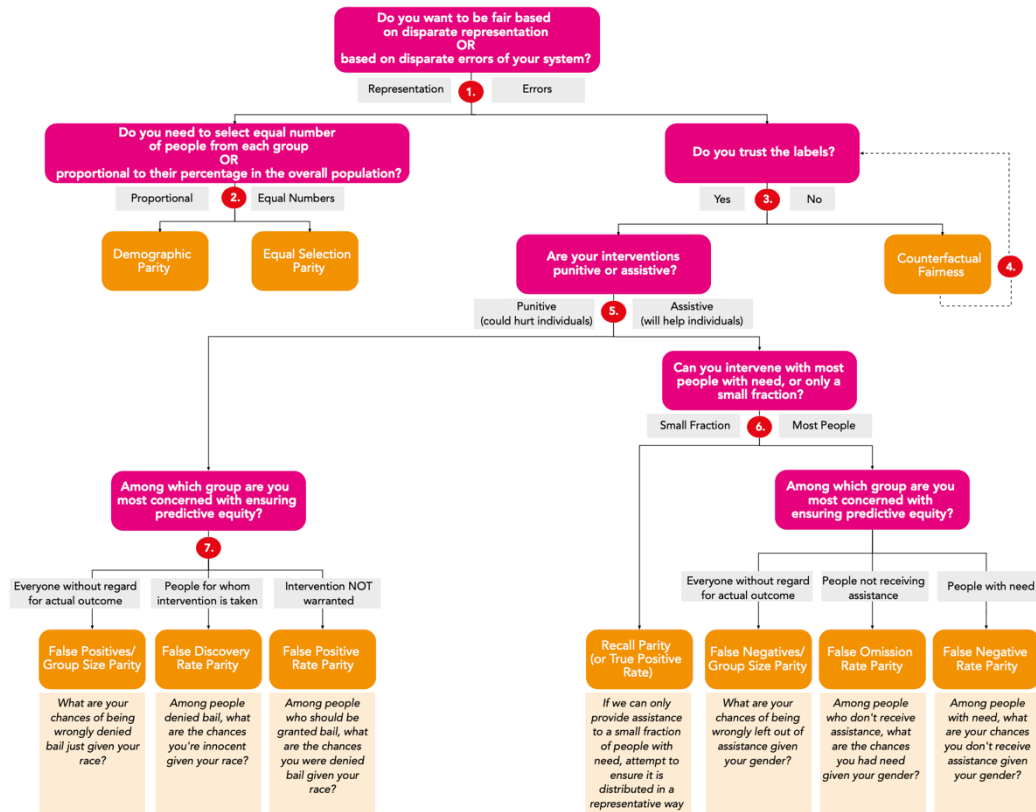
Fairness Measures: Which Measure to choose?



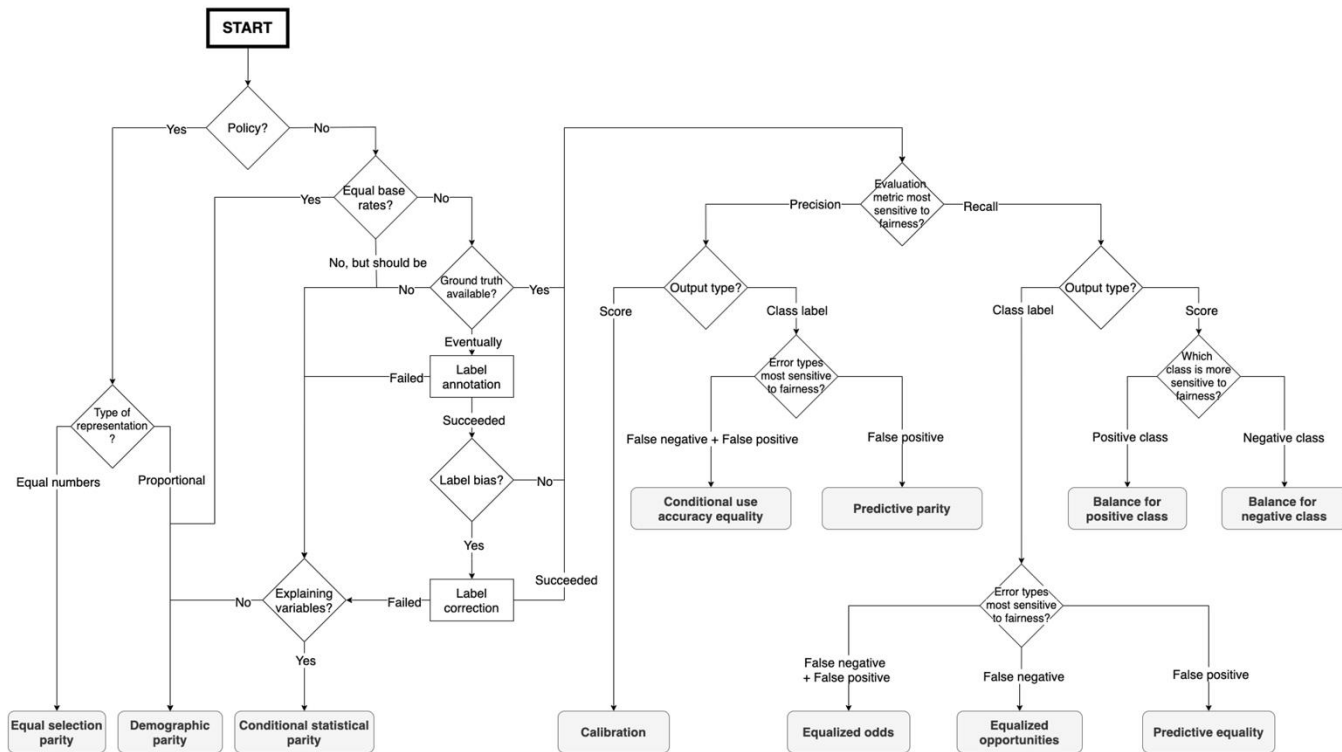
Fairness Measures: Which Measure to choose?

- One of the main challenges in fairness analysis is choosing a suitable fairness definition and corresponding measure.
- Each stakeholder might have a different understanding of fairness (e.g., COMPAS case), which can hinder the agreement on what it means to be “fair”.
- Each fairness metric has its own goals, opportunities, and constraints, and they are often conflicting.
- **So how can we approach the definition and selection of a fairness measure?**

Fairness Measures: Aequitas Fairness Tree



Fairness Measures: Fairness Compass

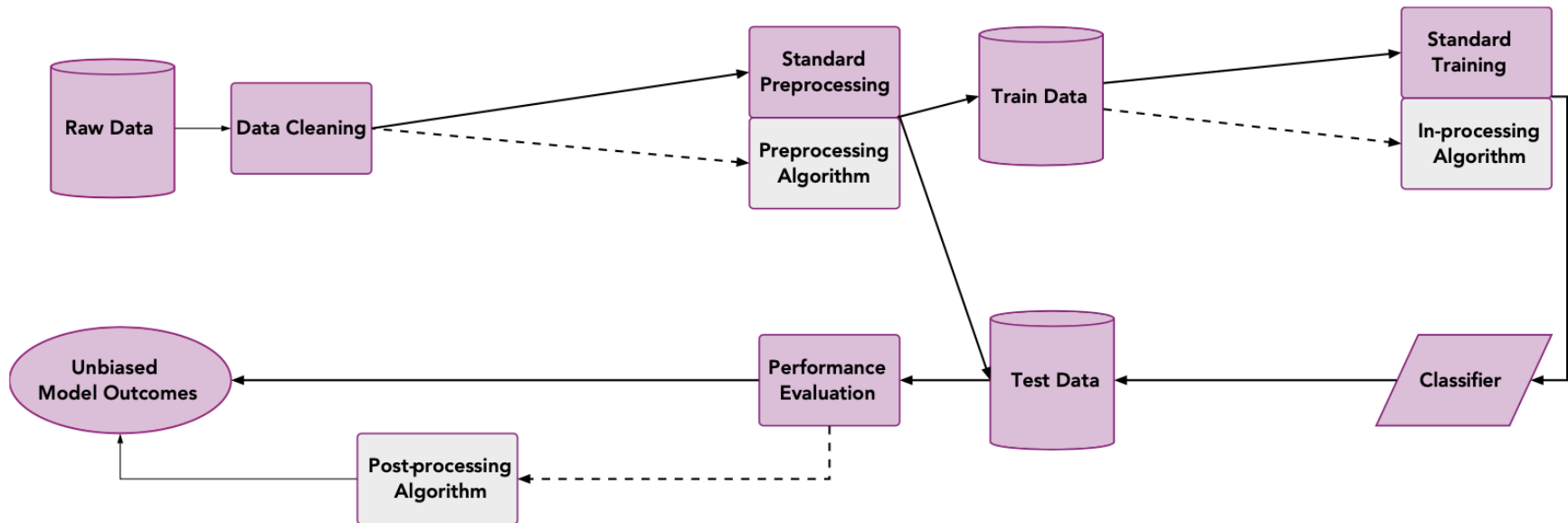


Mitigating Bias in AI

Pre-Processing, In-Processing, and Post-Processing Strategies

Bias Reduction and Mitigation Strategies

- Generally, methods for fair machine learning fall under three categories:



Bias Reduction and Mitigation Strategies: Pre-processing Techniques

- **Pre-processing techniques** address bias by **transforming the data before creating the model** in order to remove the underlying discrimination. They do not need to access or modify the model.
- These may include **modifying the labels**, the **observed data**, and the **weighting** of the features. After **de-biasing the data**, the models can be learned in the standard way.

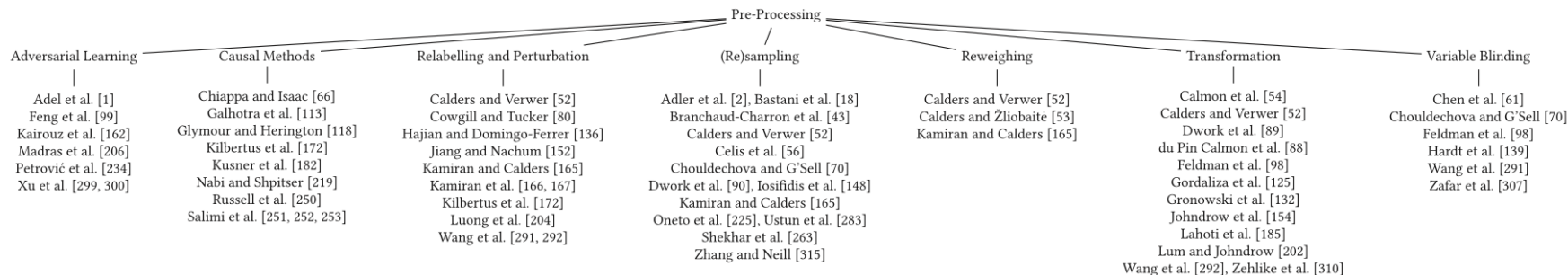


Fig. 2. Pre-processing methods.

Bias Reduction and Mitigation Strategies: In-processing Techniques

- **In-processing techniques** address bias by **modifying the learning algorithms** to remove discrimination during the training process of the models.
- This may include specifying **custom objective functions** or imposing **fairness constraints**. In-processing allows the highest flexibility to find **suitable trade-offs** between performance and fairness, but it **requires access to the model** (and is also dependent on the type of ML model).

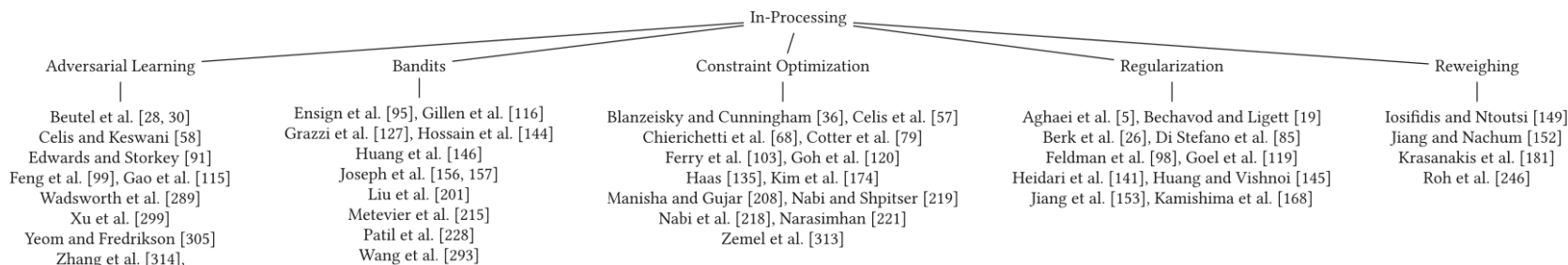


Fig. 3. In-processing methods.

Bias Reduction and Mitigation Strategies: Post-processing Techniques

- **Post-processing techniques** address bias by manipulating the output predictions in order to optimize a fairness metric.
- This is done by finding **suitable thresholds for each group** that results in equal prediction distributions. Alternatively, post-processing techniques can directly **intervene on a classification threshold** that ensures (or fosters) fair outcomes.

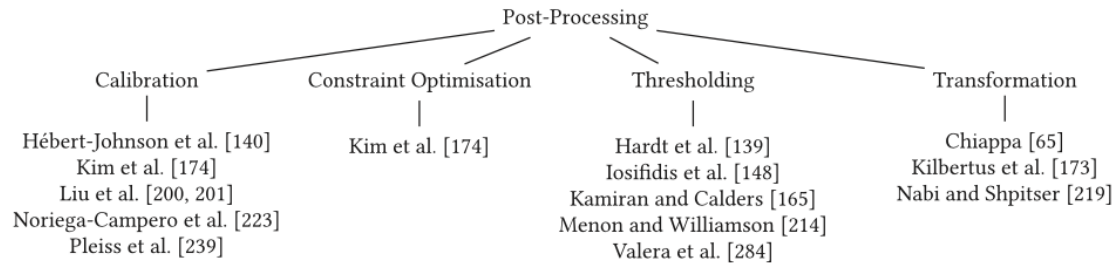
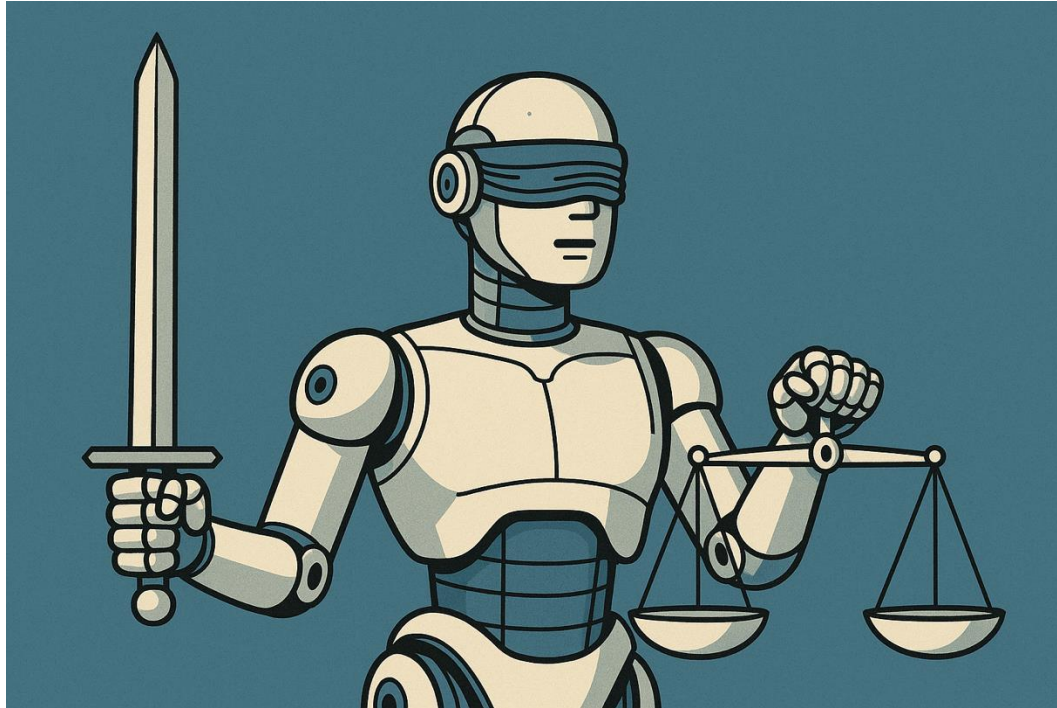


Fig. 4. Post-processing methods.

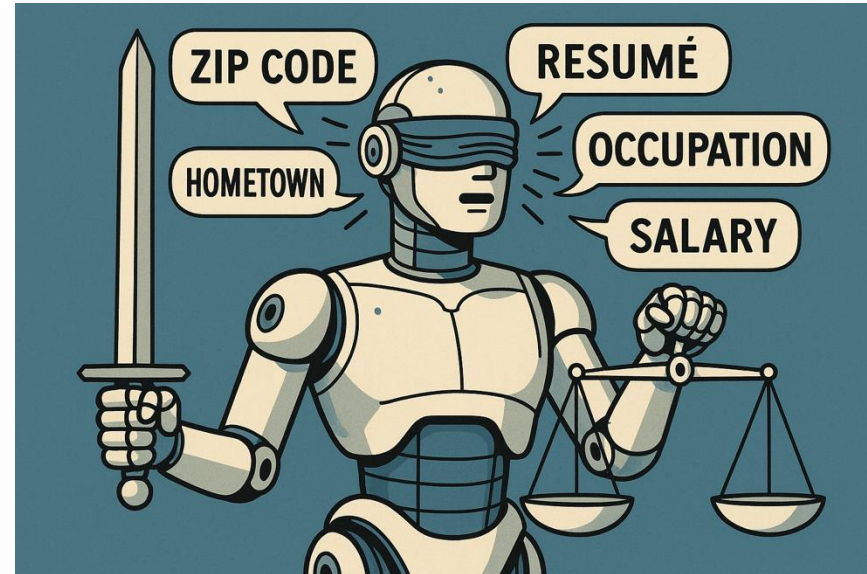
Bias Reduction and Mitigation Strategies: Fairness Through Unawareness

- If the model doesn't know the race, sex, age, etc,... **how can it discriminate?**



Bias Reduction and Mitigation Strategies: Fairness Through Unawareness

- If the **model doesn't know** the race, sex, age, etc,... **how can it discriminate?**
- **Fairness Through Unawareness:** refers to **leaving out of the model protected attributes** such as gender, race, and other characteristics deemed sensitive.
- **However!!!**
 - Other attributes that remain unprotected might still be highly correlated with the protected attributes.
 - A race/sex/etc blind model can still discriminate.
 - **Ignoring meaningful group differences does not erase inequality but might exacerbate and perpetuate it instead.**
 - **There is no fairness through unawareness.**



Bias and Fairness

Practice with Python

holisticai: Holistic AI Library

- Open-source tool to **assess and improve the trustworthiness of AI systems**. Offers a set of techniques **to measure and mitigate bias**.

```
# train a logistic regression model
model = LogisticRegression(random_state=42, max_iter=500)
model.fit(X_train_t, train_data['y'])

# make predictions
y_pred = model.predict(X_test_t)

# compute bias metrics
metrics = classification_bias_metrics(
    group_a = test_data['group_a'],
    group_b = test_data['group_b'],
    y_true = test_data['y'],
    y_pred = y_pred
)

# create a comprehensive report
bias_metrics_report(model_type='binary_classification', table_metrics=metrics)
```

	Baseline	Preprocessing Mitigator	Inprocessing Mitigator	Postprocessing Mitigator	Reference
Metric					
Statistical Parity	0.174129	0.175140	0.108639	0.062752	0
Disparate Impact	3.003852	3.031482	2.114219	1.369805	1
Four Fifths Rule	0.332906	0.329872	0.472988	0.730031	1
Cohen D	0.440582	0.443250	0.291181	0.153884	0
2SD Rule	19.193775	19.305206	12.833175	6.827009	0
Equality of Opportunity Difference	0.085004	0.081466	-0.037908	-0.159924	0
False Positive Rate Difference	0.070272	0.072136	0.030534	-0.017353	0
Average Odds Difference	0.077638	0.076801	-0.003687	-0.088638	0
Accuracy Difference	-0.104368	-0.106390	-0.110517	-0.067848	0



<https://holisticai.readthedocs.io>



`pip install holisticai`

giskard: Open-Source AI testing library

- 🦛 Open-source evaluation and testing for LLMs and ML models

 Langchain HuggingFace PyTorch Tensorflow scikit-learn any Python function

```
import giskard
from sklearn.pipeline import Pipeline

# Pipeline for the sklearn model
clf = Pipeline(...)
clf.fit(...)


# Wrap your Pandas DataFrame
dataset = giskard.Dataset(
    df=titanic_df, target="Survived"
)

# Wrap your model
model = giskard.Model(
    model=clf.predict_proba,
    model_type="classification"
)

# Scan for vulnerabilities
results = giskard.scan(model, dataset)
```

```
display(results)

# Save it to a file
results.to_html("scan_report.html")
```

 We found some potential spurious correlations between your data and the model predictions. Some data slices are highly correlated with your predictions. This happens when:

- Data leakage: one of the feature is indirectly linked with the target variable
- Overfitting: the model learns specific noisy patterns of the training data, including coincidental correlations that are not causal
- Data noise: the training set contains anomalies that are unrelated to the underlying problem (data collection, measurement biases, data preprocessing issues, etc.)

To learn more about causes and solutions, check our [guide on spurious correlation](#).

ISSUES 3 INFO

<code>`Sex` == "female"</code>	Nominal association (Theil's U) = 0.697	Prediction Survived = 'yes' for 92.67% of samples in the slice	Show details
<code>`Sex` == "male"</code>	Nominal association (Theil's U) = 0.697	Prediction Survived = 'no' for 96.28% of samples in the slice	Show details
<code>`Name` contains "mr"</code>	Nominal association (Theil's U) = 0.609	Prediction Survived = 'no' for 98.48% of samples in the slice	Show details

<https://docs.giskard.ai>`pip install giskard`

Example on: https://docs.giskard.ai/en/stable/getting_started/quickstart/quickstart_tabular.html

fairlearn: Assess and improve fairness of machine learning models

Example Notebooks

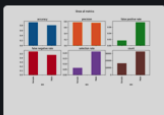
Here's a list of examples on how to use the library. We will be adding more examples soon. If you're interested in contributing to existing notebooks or adding new ones please consult the guide on [Contributing example notebooks](#).

Note

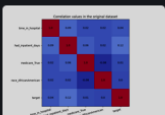
The Fairlearn API is still evolving, so if you want to run these on your local Fairlearn installation, make sure to match versions.



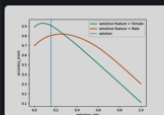
Selection rates in census dataset



MetricFrame visualizations



CorrelationRemover visualization



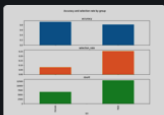
Passing pipelines to mitigation techniques



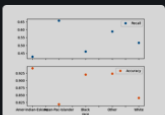
MetricFrame: Beyond Binary Classification



Making Derived Metrics



GridSearch with Census Data



Plotting Metrics with Errors



Metrics with Multiple Features



Intersectionality in Mental Health Care

- A Python package to assess and improve fairness of machine learning models.

```
metrics = {
    "accuracy": accuracy_score,
    "precision": precision_score,
    "false positive rate": false_positive_rate,
    "false negative rate": false_negative_rate,
    "selection rate": selection_rate,
    "count": count,
}

metric_frame = MetricFrame(
    metrics=metrics, y_true=y_true, y_pred=y_pred, sensitive_features=sex
)

metric_frame.by_group.plot.bar(
    subplots=True,
    layout=[3, 3],
    legend=False,
    figsize=[12, 8],
    title="Show all metrics",
)
```



<https://fairlearn.org/v0.10/quickstart.html>



pip install fairlearn

aequitas: Bias and Fairness Audit Toolkit for Machine Learning

Aequitas
Bias & Fairness Audit

Home Code About

Bias and Fairness Audit Toolkit

The Bias Report is powered by [Aequitas](#), an open-source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools.



See an [example report](#) on COMPAS risk assessment scores.

Or try out the audit tool using your own data or one of our sample data sets.

Get Started!



<https://github.com/dssg/aequitas>



pip install aequitas

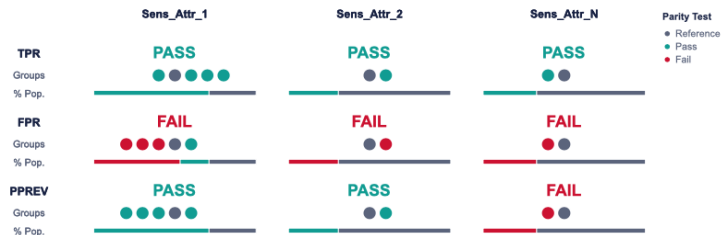
- Open-source bias auditing and fair machine learning toolkit for data scientists.

```
from aequitas import Audit
```

```
audit = Audit(df)
```

To obtain a summary of the bias audit, run:

```
# Select the fairness metric of interest for your dataset
audit.summary_plot(["tpr", "fpr", "pprev"])
```



For a group to pass the parity test its disparity to the reference group cannot exceed the fairness threshold (1.25).
 An attribute passes the parity test for a given metric if all its groups pass the test.

aif360: AI Fairness Toolkit

- Open-source library to detect and mitigate bias in machine learning models.

AI Fairness 360 (AIF360)

Continuous Integration failing docs passing pypi package 0.6.1 CRAN not published

The AI Fairness 360 toolkit is an extensible open-source library containing techniques developed by the research community to help detect and mitigate bias in machine learning models throughout the AI application lifecycle. AI Fairness 360 package is available in both Python and R.

The AI Fairness 360 package includes

1. a comprehensive set of metrics for datasets and models to test for biases,
2. explanations for these metrics, and
3. algorithms to mitigate bias in datasets and models. It is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

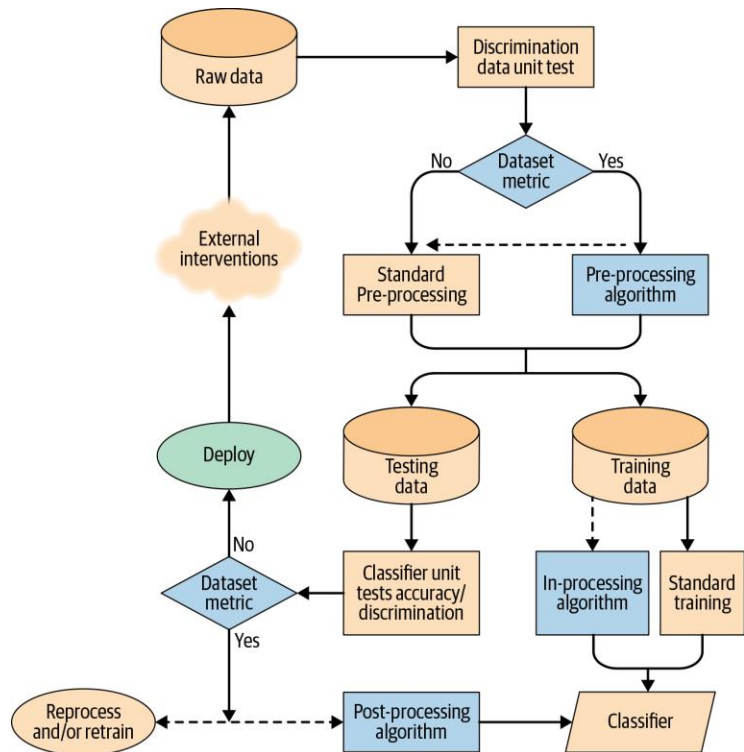


<https://github.com/Trusted-AI/AIF360>



`pip install aif360`

Book Tutorial: <https://krvarshney.github.io/pubs/MahoneyVH2020.pdf>



Responsible AI Dashboard: Customized, end-to-end RAI experience

- A collection of integrated tools and functionalities to help operationalize Responsible AI in practice. In particular, you can explore the **Fairness Dashboard**, which allows to understand model's unfairness issues using various group-fairness metrics across sensitive features (**powered by fairlearn**).

Welcome to the Fairness dashboard

The fairness dashboard enables you to assess tradeoffs between performance and fairness of your models



01 Sensitive features 02 Performance metrics 03 Fairness metrics

How do you want to measure fairness?

Fairness metrics quantify variation of your model's behavior across selected features. There are several kinds of fairness metrics that are based on a variety of performance metrics. They either capture the difference or ratio between the extreme values across the groups, or simply the worst value of any group.

Data statistics

2 sensitive features
9769 instances

> Accuracy / error rate (6)

▼ Demographic Parity / Selection rate (2)

- ☒ **Demographic parity difference** The maximum difference in selection rate, that is the fraction with predicted label ...
- ☐ Demographic parity ratio The minimum ratio of selection rates, that is the fraction with predicted label 1, be...

01 Sensitive features

02 Performance metrics

03 Fairness metrics

Along which features would you like to evaluate your model's fairness?

Fairness is evaluated in terms of disparities in your model's behavior. We will split your data according to values of each selected feature, and evaluate how your model's performance metric and predictions differ across these splits.

Data statistics

2 sensitive features
9769 instances

Sensitive features	Subgroups
<input checked="" type="checkbox"/> sex This feature has 2 unique values	Male Female
<input type="checkbox"/> race This feature has 5 unique values	White Black Asian-Pac-Islander Other Amer-Indian-Eskimo

Next



<https://github.com/microsoft/responsible-ai-toolbox>

`pip install raiwidgets`

RAI Dashboard: <https://responsibleaitoolbox.ai/>

Datasets for Fairness-Aware ML

Exploring bias and fairness in publicly available data

Datasets for Fairness-Aware Machine Learning

- Quy *et al.* present a benchmark of real-world tabular datasets from fairness experiments.
- They use a Bayesian network to identify the relationship between protected attributes and the outcome, and further explore bias in data using exploratory data analysis.
- Datasets are categorized by application domain:
 - **Financial datasets (6):** *adult, kdd-census-income, german-credit, dutch-census, bank-marketing, credit-card-clients*
 - **Criminological datasets (3):** *compas-recid, compas-viol-recid, communities-and-crime*
 - **Healthcare and Social datasets (2):** *diabetes, ricci*
 - **Educational datasets (2):** *student-mat, student-port, oulad, law-school*

Datasets for Fairness-Aware Machine Learning

TABLE 1 Overview of real-world datasets for fairness

Dataset	#Instances	#Instances (cleaned)	#Attributes (cat./bin./ num.)	Class	Domain	Class ratio (+:−)	Protected attributes	Target class	Collection period	Collection location
Adult	48,842	45,222	7/2/6	Binary	Finance	1:3.03	Sex, race, age	Income	1994	USA
KDD Census-Income	299,285	284,556	32/2/7	Binary	Finance	1:15.30	Sex, race	Income	1994–1995	USA
German credit	1000	1000	13/1/7	Binary	Finance	2.33:1	Sex, age	Credit score	1973–1975	Germany
Dutch census	60,420	60,420	10/2/0	Binary	Finance	1:1.10	Sex	Occupation	2001	The Netherlands
Bank marketing	45,211	45,211	6/4/7	Binary	Finance	1:7.55	Age, marital	Deposit subscription	2008–2013	Portugal
Credit card clients	30,000	30,000	8/2/14	Binary	Finance	1:3.52	Sex, marriage, education	Default payment	2005	Taiwan
COMPAS recid.	7214	6172	31/6/14	Binary	Criminology	1:1.20	Race, sex	Two-year recidivism	2013–2014	USA
COMPAS viol. recid.	4743	4020	31/6/14	Binary	Criminology	1:5.17	Race, sex	Two-year violent recid.	2013–2014	USA
Communities and Crime	1994	1994	4/0/123	Multi	Criminology	—	Black	Violent crimes rate	1995	USA
Diabetes	101,766	45,715	33/7/10	Binary	Healthcare	1:3.13	Gender	Readmit in 30 days	1999–2008	USA
Ricci	118	118	0/3/3	Binary	Society	1:1.11	Race	Promotion	2003	USA
Student—Mathematics	649	649	4/13/16	Binary	Education	1:2.04	Sex, age	Final grade	2005–2006	Portugal
Student—Portuguese	649	649	4/13/16	Binary	Education	1:5.49	Sex, age	Final grade	2005–2006	Portugal
OULAD	32,593	21,562	7/2/3	Multi	Education	—	Gender	Outcome	2013–2014	England
Law School	20,798	20,798	3/3/6	Binary	Education	8.07:1	Male, race	Pass the bar exam	1991	USA

Abbreviations: COMPAS, Correctional Offender Management Profiling for Alternative Sanctions; OULAD, Open University Learning Analytics dataset.

References and Further Reading

- Muhammad, S. (2022). The Fairness Handbook.
- Foster, Ian, et al., eds. Big data and social science: Data science methods and tools for research and practice. *CRC Press*, 2020.
- Suresh, Harini, and John Gutttag. "A framework for understanding sources of harm throughout the machine learning life cycle." *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2021.
- Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *ACM computing surveys (CSUR)* 54.6 (2021): 1-35.
- Caton and Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* 56, 7, 166-
- Le Quy, Tai, et al. "A survey on datasets for fairness-aware machine learning." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12.3 (2022): e1452.
- Verma, Sahil, and Julia Rubin. "Fairness definitions explained." *Proceedings of the international workshop on software fairness*. 2018.
- Ruf, Boris, and Marcin Detyniecki. "Towards the right kind of fairness in AI." *arXiv preprint arXiv:2102.08453* (2021).
- Ntoutsis, Eirini, et al. "Bias in data-driven artificial intelligence systems -- An introductory survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.3 (2020): e1356.

Tutorial

T04: Bias and Fairness

Artificial Intelligence and Society

Module 04: Bias and Fairness

Miriam Seoane Santos

LIAAD, INESC TEC, FCUP, University of Porto

miriam.santos@fc.up.pt