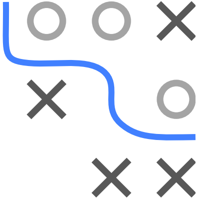
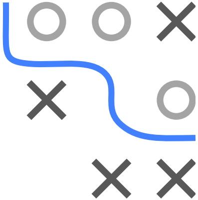
# ALGORITHMIC FAIRNESS



- Machine learning (ML) based systems increasingly permeate society
- Models can replicate existing injustices or introduce new ones
- Automated decisions can disproportionately harm vulnerable individuals

# ALGORITHMIC FAIRNESS



## Medicine

Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

www.pnas.org/content/117/23/12592

## Criminal Justice

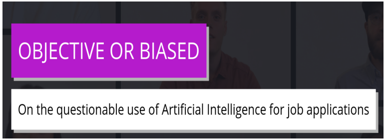
Machine Bias

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

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

## Hiring




https://interaktiv.br.de/ki-bewerbung/en/

## Search Results

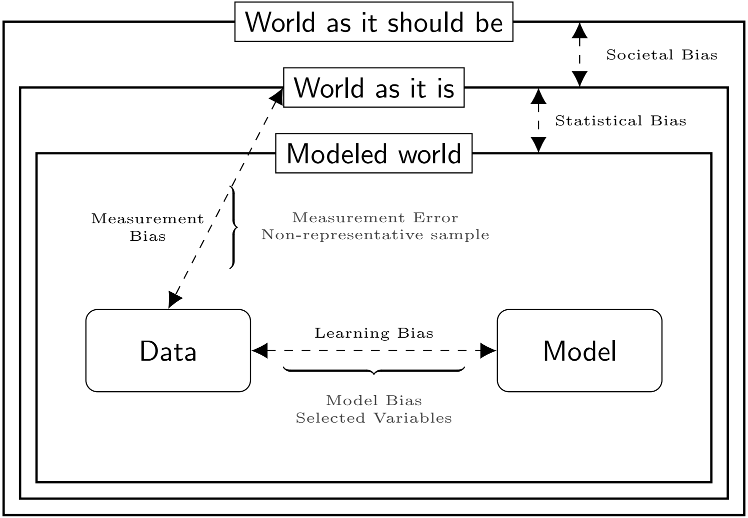
IDEAS • TECH BIAS

Google Has a Striking History of Bias Against Black Girls

BY SAFIYA NOBLE

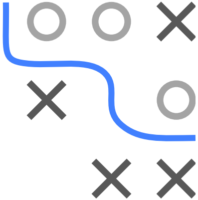
https://time.com/5209144/google-search-engine-algorithm-bias-racism/

# SOURCES OF BIAS

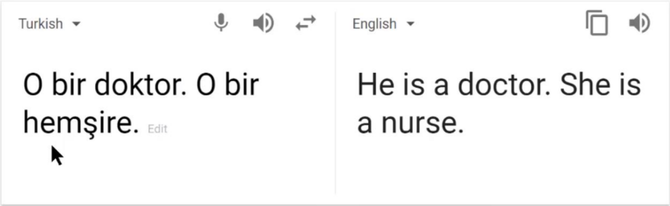



Adapted from S. Mitchell et al., Algorithmic fairness: Choices, assumptions, and definitions, 2021

# HISTORICAL BIAS



- Historical data often contains biases, e.g. under-representation of minority groups
- Models can pick up existing biases
- As a result, biases are perpetuated into the future

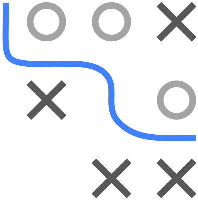



Twitter: math_rachel

01.05.2019

# REPRESENTATION BIAS

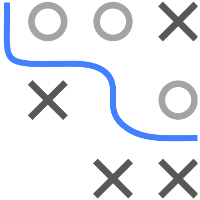
- Over- or under-representation of specific sub-population can lead to models that only predict well for majority groups
- Models need to be evaluated across a representative sample of the target population
- Example: We can only know if a person paid back a loan if we gave out a loan in the first place



| | TYPE I | TYPE II | TYPE III | TYPE IV | TYPE V | TYPE VI |
|---|---|---|---|---|---|---|
| | 1.7% | 1.1% | 3.3% | 0% | 23.2% | 25.0% |
| IBM | 5.1% | 7.4% | 8.2% | 8.3% | 33.3% | **46.8%** |
| FACE++ | 11.9% | 9.7% | 8.2% | 13.9% | 32.4% | **46.5%** |

gendershades.org

# OTHER SOURCES OF BIAS



- **Measurement Bias** Difference in how a given variable is measured in different sub-populations
  - Increased policing in some post codes lead to more prior arrests
  - Better data quality between different hospitals
- **Model Bias** Biases introduced during modelling, e.g. due to under-specified models
  - Models make more errors for darker skin tones due to insufficient data
  - Models pick up spurious correlations in the data
- **Feedback Loops** Model decisions shape data collected in the future
  - Lead to representation bias if e.g. sub-populations are systematically excluded
  - People and ML systems ’pick up’ miss-representation from search engines.

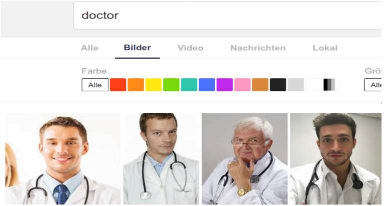
Mehrabi et al., A Survey on Bias and Fairness in Machine Learning, 2020

# TYPES OF HARMS

If not accounted for, *biases* can lead to several **harms**

- **Allocation**: A ressource is allocated unevenly across individuals
- **Quality-of-service**: Systems fail disproportionately for certain groups of individuals.
- **Stereotyping**: Systems re-inforce existing stereotypes
- **Denigration**: Systems are offensive towards individuals
- **Representation**: Under- or overrepresentation of certain groups

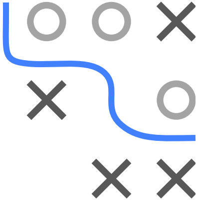

Twitter: jackyalcine 29.06.2015




google.com search for doctor (May, 2021)

H. Weerts, An introduction to algorithmic fairness, 2021

# AUDITING MODELS FOR POTENTIAL HARMS



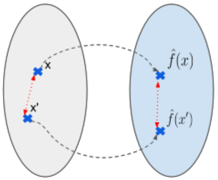
For a more formal treatment, we introduce additional notation:

- **Protected attribute:** A protected *class* or attribute w.r.t which models should be fair.
  - We denote this protected attribute $A$ with $\mathbf{a}$.
  - For simplicity, we assume that $\mathbf{a}^{(i)} \in \mathcal{A} = \{0, 1\}$ is a binary variable.
- **Decision space:** To differentiate between a model's prediction $\hat{f}(\mathbf{x})$ and a decision derived from this prediction, we denote the decision with $\mathbf{d}$. For simplicity, we assume $\mathbf{d}^{(i)} \in \delta = \{0, 1\}$
- This notation can be extended to multi-class or regression outcomes as well as more complex protected attributes, e.g. that account for non-binary protected classes or *intersectional notions*, e.g. $\texttt{race} \wedge \texttt{gender}$.

# MATHEMATICAL NOTIONS OF BIAS - OVERVIEW

**Individual Fairness**

Similar individuals should be treated similarly

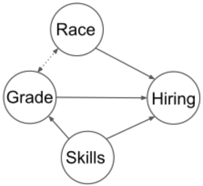



**Statistical (Group) Fairness**

Define fairness as an average disparity across protected classes (e.g. race, gender, ...)

**Causal Fairness**

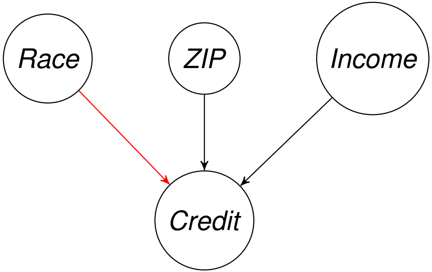
Fairness notions should take causal relationships in the data into account



# NO FAIRNESS THROUGH UNAWARENESS

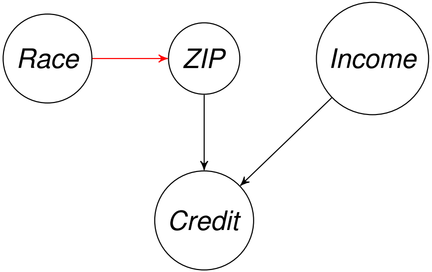
A naive proposal to reduce harms from ML models is to simply remove the protected attribute. **But:** It's not that simple - models can pick up the information through other variables!

**Direct Discrimination**



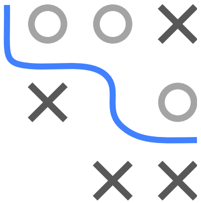

$\rightarrow$ The model directly uses race as a feature.

**Indirect Discrimination**




$\rightarrow$ The model picks up information about the race through the proxy-variable ZIP-code.

# GROUP FAIRNESS DEFINITIONS



Several fairness definitions based on differences between protected groups have been proposed.

- **Statistical Parity**: The chance to get the favourable outcome is equal across two groups. This is also called *demographic parity*.

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

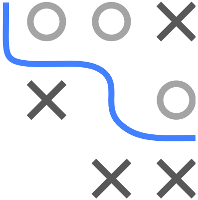
- **Equalized Opportunity**: The chance to *correctly* be assigned the favourable outcome is independent of the protected attribute.

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

- **Accuracy Parity**: The accuracy is equal in both groups.

$$P(\hat{Y} = 1 | A = 0, Y = 1) + P(\hat{Y} = 0 | A = 0, Y = 0) =$$
$$P(\hat{Y} = 1 | A = 1, Y = 1) + P(\hat{Y} = 0 | A = 1, Y = 0)$$

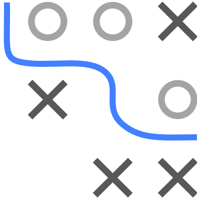
# PERSPECTIVE: BASED ON PREDICTED OUTCOME



- Statistical parity requires equality in the predicted outcome. E.g. hire candidates **independent** of qualification.
- If the underlying qualifications are not distributed equally across groups, we need to sacrifice *utility* to achieve statistical parity.


$\rightarrow$ Enforcing equal positive rates might require hiring unqualified candidates.
**Danger:** If the bias comes from the real world (e.g. societal bias), enforcing statistical parity can also lead to adverse effects in the long term.

# PERSPECTIVE: BASED ON TRUE & PREDICTED OUTCOME



- Other fairness notions require equality of some error notions, e.g. false positive rates. E.g. hire *qualified* candidates at equal rates across groups.
- Error based notions are often more intuitive and easy to communicate.
- Can help to idenitify representation or model bias.
- Error based notions do not account for systemic injustices in the world – if e.g. labels are biased, we can still be *fair* according to error-based notions.

# REMINDER: CONFUSION MATRIX

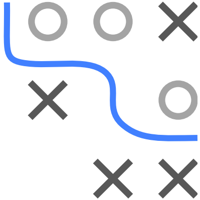
The confusion matrix is a $2 \times 2$ contingency table of predictions $\hat{y}$ and true labels $y$. Several evaluation metrics can be derived from a confusion matrix:

| | Actual Positive | Actual Negative | |
|---|---|---|---|
| Predicted Positive | **True positive (TP)** | **False positive (FP, Type I error)** | Precision (P) = Positive predictive value $= \frac{\#TP}{\#TP + \#FP}$ |
| Predicted Negative | **False negative (FN, Type II error)** | **True negative (TN)** | Negative predictive value $= \frac{\#TN}{\#FN + \#TN}$ |
| | Sensitivity = Recall (R) = True positive rate (tpr) $= \frac{\#TP}{\#TP + \#FN}$ | Specificity = True negative rate (tnr) $= \frac{\#TN}{\#FP + \#TN}$ | Accuracy $= \frac{\#TP + \#TN}{n}$ Error rate $= \frac{\#FN + \#FP}{n}$ |

$\rightarrow$ Many fairness metrics can be expressed as entries of the confusion matrix

# FAIRNESS TENSOR

We can represent labels & predictions as a *fairness tensor* (Kim et al., 2020). Fairness tensors are 3-dimensional, stacked confusion matrices:

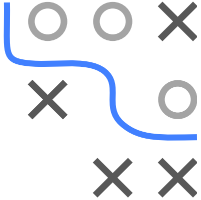


$$Z = \left[ \begin{bmatrix} TP_1 & FP_1 \\ FN_1 & TN_1 \end{bmatrix}^{A=1}, \begin{bmatrix} TP_0 & FP_0 \\ FN_0 & TN_0 \end{bmatrix}^{A=0} \right]$$

For $z = (TP_1, FN_1, FP_1, TN_1, TP_0, FN_0, FP_0, TN_0)^T / N$, we can express a large variety of fairness metrics as linear $\phi(x) = A \cdot z$ or quadratic functions $\phi(x) = z^T \cdot B \cdot z$ by choosing an appropriate matrix $A$ or $B$.

**Example:**
We choose $A = (N_1, 0, N_1, 0, N_0, 0, N_0, 0)/N$, where $N_a$ is the sum of entries in the confusion matrix for protected group $a$. We can now express **statistical parity** as $A \cdot z = 0$.

# INCOMPATIBILITY OF FAIRNESS METRICS

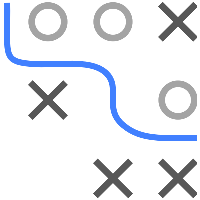


- Some fairness metrics cannot be jointly satisfied.
- E.g. simultaneously satisfying equal *TPR*, *FPR*, and *FNR*.
- Question: how can we show the above point formally?
- Answer:
  - Using the fairness tensor $z$ and $A_{TPR}, A_{FPR}, A_{FNR}$ to encode the fairness metrics.
  - Making the fairness metrics compatible needs $z$ to fufill

$$\begin{bmatrix} A_{TPR} \\ A_{FPR} \\ A_{FNR} \end{bmatrix} \cdot z = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

  - If no valid solution $z$ exists, the metrics are incompatible.

# FAIRNESS METRICS - CLOSING THOUGHTS

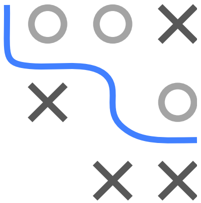


- Statistical group fairnes metrics require translating ethical considerations of what is *fair* into mathematical formulas.
- To draw meaningful conclusions, we need to evaluate fairness metrics on a **representative** data set.
- Fairness metrics reduce a wide variety of important considerations into a single number – they are not designed to guarantee that a system is fair.
- Incompatibility between fairness metrics implies that we might need trade-offs between fairness metrics.

# PREVENTING & MITIGATING HARMS - DOCUMENTATION

- Idea: prevent harms of ML models by improving documentations of models & datasets.
- Motivation: usage of datasets or models outside of their intended use can often lead to harm, even if the models are carefully validated.
- **Dataset documentation** Includes information on the dataset, sampling mechanisms and intended use.
- **Model documentation** Includes information about the model, used data and hyperparameters.
- **Fairness reports** Include information about performed fairness audits.

# PREVENTING & MITIGATING HARMS - BIAS MITIGATION

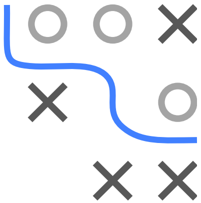


Several *bias mitigation techniques* have been proposed:

- **Pre-processing**: Transform data to make subsequently trained models fairer.
- **In-processing**: Learn a model that directly incorporates fairness constraints.
- **Post-processing**: Adapt model predictions to satisfy fairness constraints.

**Example:** Re-weighing (Kamiran, 2012) proposes to use sample weights that are inverse to the frequency of labels and predictions in the data.

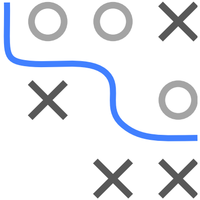
# PREVENTING & MITIGATING HARMS - RECOURSE



Fair treatment of individuals subject to a decision making systems decisions can often not only be achieved solely through algorithmic means but requires recourse, accountability & interpretability.

- **Accountability**: Automated systems will make errors - developers need to ensure that humans responsible for addressing such errors exist and have the means to address such errors.
- **Interpretability**: Interpretability techniques can help to identify possible problems in the data or the model, e.g. spurious correlations picked up by the model.
- **Recourse**: Individuals subject to automated decisions should have access to an explanation on how the decision was made and what steps can be taken to address unfavourable predictions.

# FURTHER CONSIDERATIONS



- **Intersectionality**: Fairness considerations should often hold across intersectional groups, e.g. *race* $\wedge$ *gender*.
- **Intervention design**: Instead of ensuring a given intervention is fair, it can often be helpful to consider the intervention we wish to deploy.
  **Example:** Instead of penalizing defendants for not showing up to court, provide them with means of transportation.
- **Stakeholder participation:** Developing ML models should take the perspective of all stakeholders such as the individuals affected by the intervention and advocacy groups.
- **Long-term perspective:** Existing metrics only consider the short-term and do not take its long-term impact into account. This might lead to adverse effects in the long-term.

# RESOURCES


- Fairness and Machine Learning - Limitations and Opportunities, Barocas et al., 2019
- Algorithmic Fairness: Choices, Assumptions, and Definitions, Mitchell et al., 2021
- A Survey on Bias and Fairness in Machine Learning, Mehrabi et al., 2020
- An Introduction to Algorithmic Fairness, H.J.P Weerts, 2021
- FACT: A Diagnostic for Group Fairness Trade-offs, Kim et al., 2020
- Data preprocessing techniques for classification without discrimination, Kamiran et al., 2012
- Fairness Through Awareness, Dwork et al., 2012