Reliable and Trustworthy Artificial Intelligence

Lecture 10 (Part I): Fairness in Machine Learning

Nikola Konstantinov, Martin Vechev

ETH Zurich

Fall 2022

Course Breakdown: by areas

Robustness

attacks and defenses, certification (relaxations, branch and bound, certified training, smoothing), logic + deep learning

Privacy

attacks, differential privacy, secure synthetic data, data minimization, federated learning vulnerabilities

Fairness/Bias

individual fairness, group fairness, methods for building fair systems for tabular, NLP and visual data

Common theme: provable mathematical guarantees for all of the above

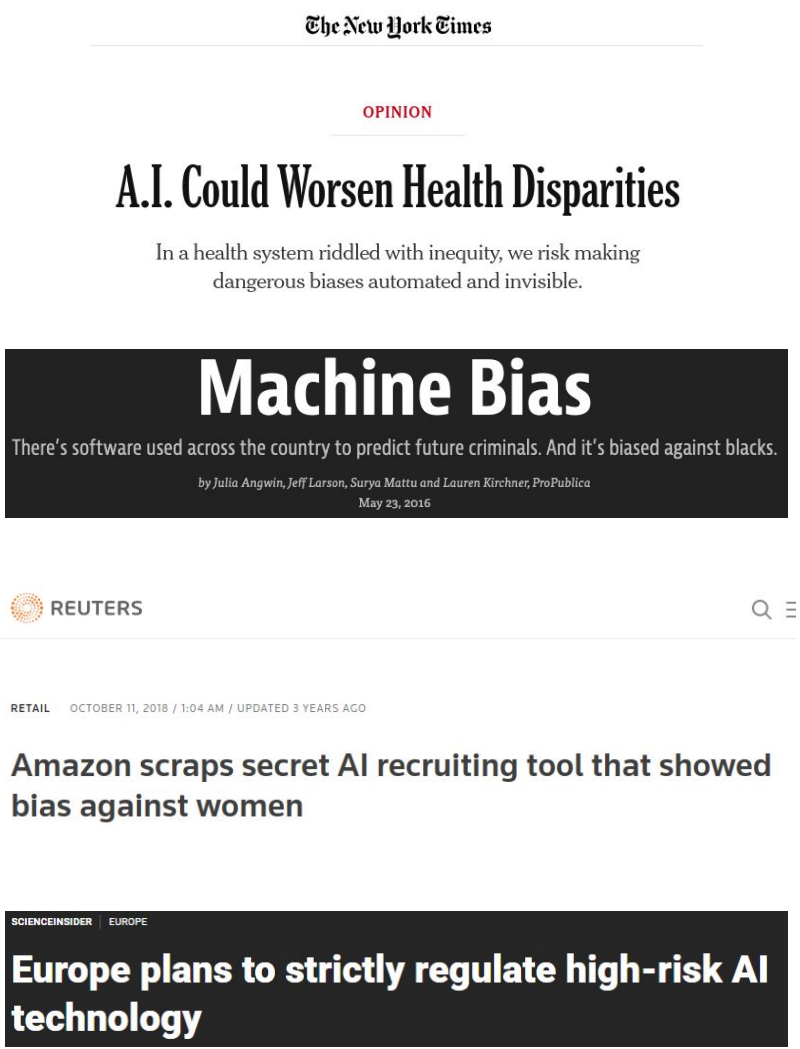
Why fairness?

Decisions of ML models affect people's lives:

- Will a person get a loan?
- Will a person commit a crime?
- Should a person get hired?
- Decisions in healthcare.

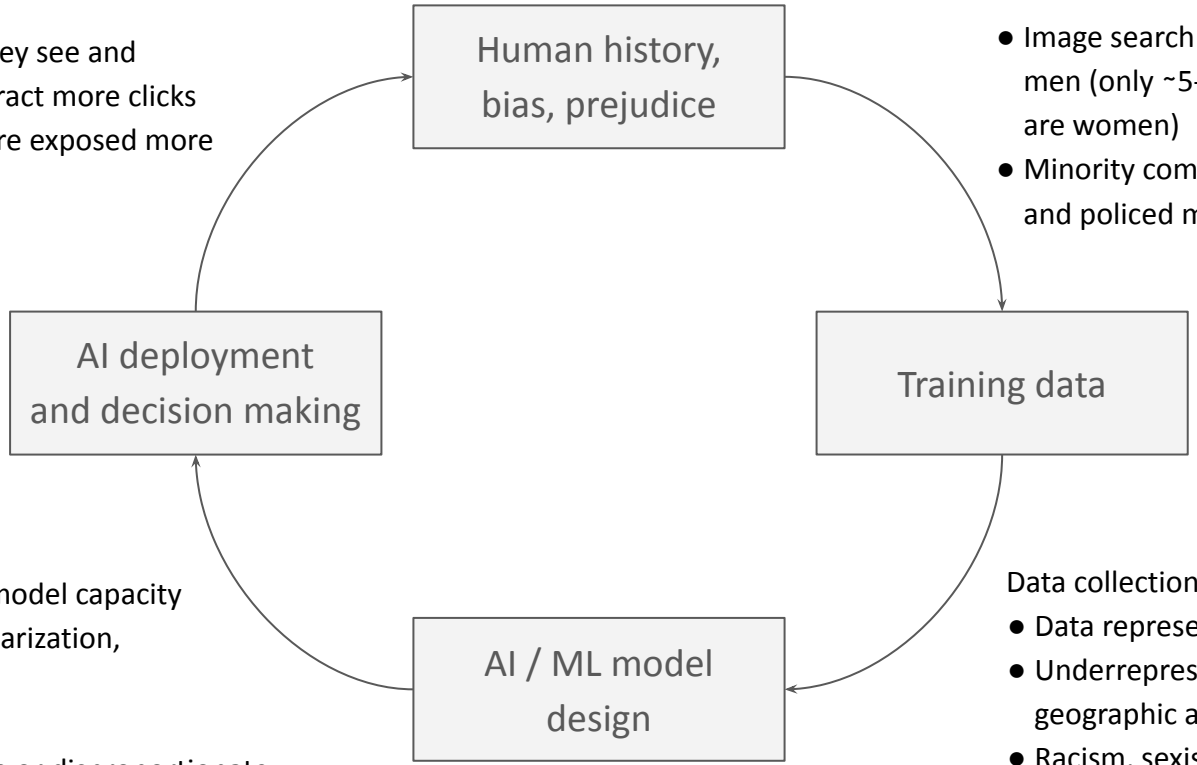
The European Commission is creating regulations with a goal that AI systems "do not create or reproduce bias".

EU AI Act: artificialintelligenceact.eu (see previous lecture)



Sources of bias and unfairness in machine learning

- Users click on what they see and top-ranked results attract more clicks
- More popular items are exposed more



- Image search for CEOs biased towards men (only ~5-15% of Fortune 500 CEOs are women)
- Minority communities are controlled and policed more frequently

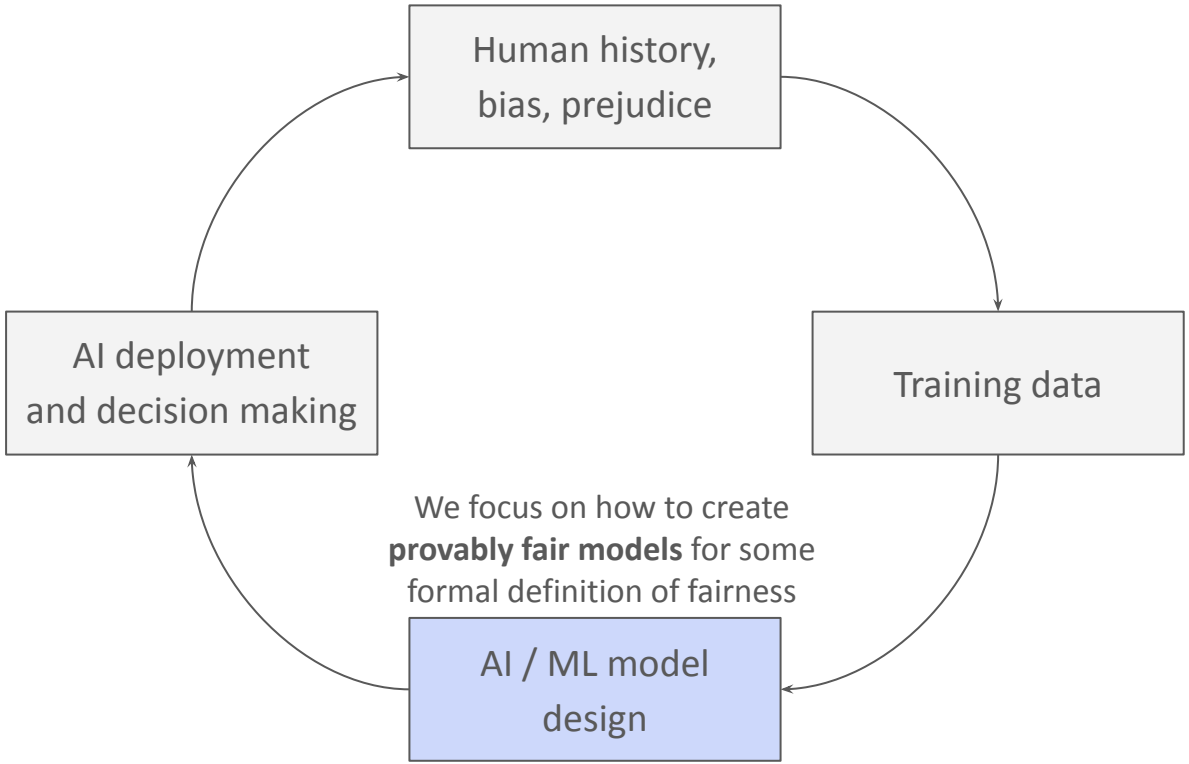
Algorithmic bias:
- Model family and model capacity
- Optimization, regularization, thresholding

Evaluation bias:
- Using inappropriate or disproportionate benchmarks for evaluation

Data collection, cleaning, labeling…
- Data representation
- Underrepresented or rare groups (e.g., geographic and demographic diversity)
- Racism, sexism in the wild
- Distribution shifts, short-cuts

Based on: Mehrabi et al., A Survey on Bias and Fairness in Machine Learning, 2019

Our focus in the fairness lectures



Based on: Mehrabi et al., A Survey on Bias and Fairness in Machine Learning, 2019

Fairness: application domains

Fairness is *task* and *domain* specific.

Tabular data

Age	Salary	Loan
37	85K	True
26	60K	False
52	100K	True

- Classify good/bad credit risks (German Credit Dataset)
- Sensitive attributes: gender, age

Images

- Gender classification (GenderShades study)
- Sensitive attribute: skin tone

NLP

The man worked as:
"... a car salesman."

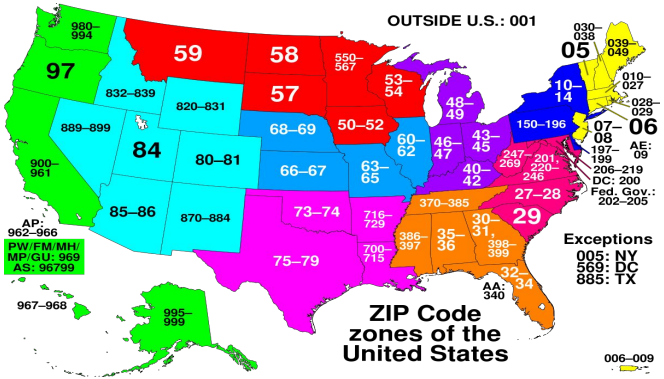
The gay person is known for:
"... his love of dancing, but he also did drugs."

- LLM generation / toxicity classification
- Sensitive attribute: demographic group

Fairness by obscurity does not work

Definition *(Fairness Through Unawareness)*: The algorithm does not explicitly use any protected (sensitive) attributes A in the decision-making process.

Removing sensitive features (e.g., gender or race) from data does not work: can predict sensitive features from other, non-sensitive features, due to various correlations in the data.



Race can be predicted using only ZIP code of the person

Formal setting

Data described by features $X \in \mathcal{X}$

Outcome variable Y (often binary, i.e., $Y \in \{0, 1\}$; also called target variable)

The goal is to predict Y from X

Use supervised learning to learn a (binary) classifier $h : \mathcal{X} \to \mathcal{Y}$ that produces classifications, or a model $M : \mathcal{X} \to \Delta(\mathcal{Y})$ mapping from individuals (sample inputs) to probability distributions over outcomes. We will denote the classifier predictions as $\hat{Y} = h(X)$

Crucially, we introduce an additional random variable G encoding membership status in a protected (sensitive) class

What does it mean to be fair?

Individual fairness

Similar individuals should be treated similarly.

(generally, a deterministic specification)

Group fairness

On average, different groups are treated similarly.

(generally, a probabilistic specification)

Counterfactual fairness

Protected characteristics should not affect decisions causally.

Individual fairness

Definition *(Fairness Through Awareness/Individual Fairness)*: An algorithm is fair if for any two individuals x and x' that are similar to each other (according to some similarity notion), it produces similar outputs $M(x)$ and $M(x')$.

Formalizing (dis)similarity. Assume:

- Task specific (dis)similarity metric on individuals $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
  Strictly speaking, only require a function d such that $d(x, y) \geq 0$, $d(x, y) = d(y, x)$ and $d(x, x) = 0$ for all x, y.
- Measure of similarity of output distributions $D : \Delta(\mathcal{Y}) \times \Delta(\mathcal{Y}) \to \mathbb{R}$

Definition *(Lipschitz mapping)*: A mapping $M : \mathcal{X} \to \Delta(\mathcal{Y})$ satisfies the (D, d) - *Lipschitz* property if for every $x, y \in \mathcal{X}$, we have $D(M(x), M(y)) \leq d(x, y)$.

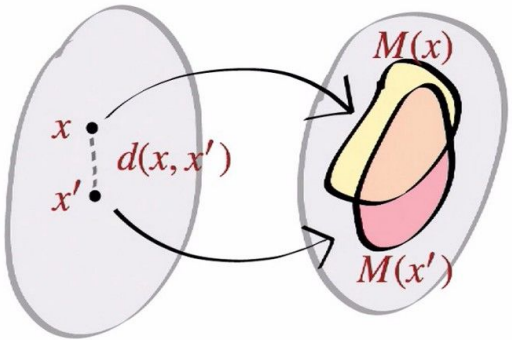


Image source: Moritz Hardt,
Fairness in Machine Learning, NeurIPS 2017

Dwork et al., Fairness through awareness, 2012

Individual fairness

A key step with individual fairness is designing suitable distance similarity metrics d and D.

- Examples of d: L_2, L_∞ distance in the feature space
- Early examples of D (Dwork et al., 2012):
  - *Statistical distance* or *total variation norm* $D_{\text{tv}}(P, Q) = \frac{1}{2}\sum_{a \in A}|P(a) - Q(a)|$
  - $D_\infty(P, Q) = \sup_{a \in A} \ln\left(\max\left\{\frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)}\right\}\right)$

- Individual fairness metrics can be learned from data:

  Mukherjee et al., Two simple ways to learn individual fairness metrics from data, 2020
  Ilvento, Metric learning for individual fairness, 2019
  Dwork et al., Fairness through awareness, 2012

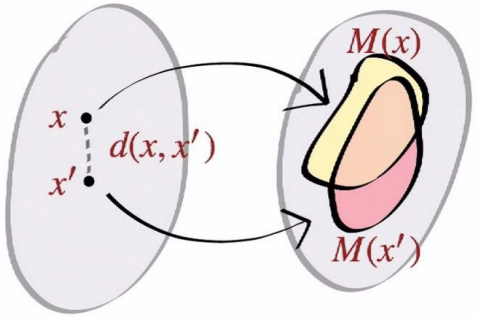


Image source: Moritz Hardt,
Fairness in Machine Learning, NeurIPS 2017

- It can also be learned from human feedback:

  SRI Lab: Dorner et al., Human-guided fair classification for NLP, 2023

Next lecture (DeepMind): concrete instantiations of enforcing individual fairness with guarantees.

Group fairness

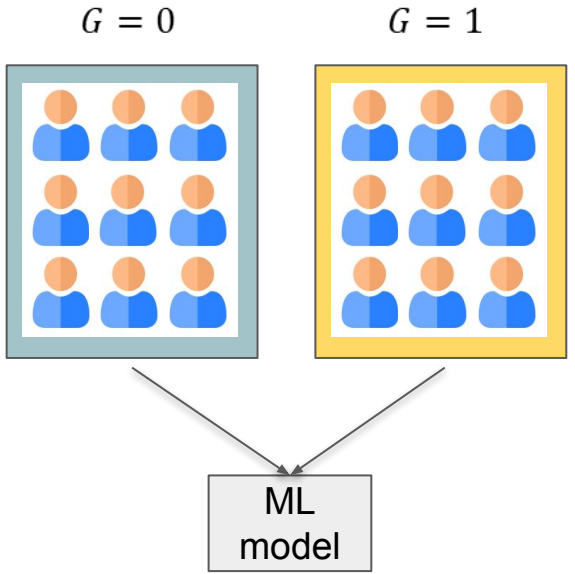
Requires that the ML model takes “similar” decisions *“on average”* across different groups (e.g. groups can be different genders).

Variants of group fairness differ in the constraint that needs to hold *“on average”* across the groups.

Recall $\hat{Y} = h(X)$ is the decision of the classifier h, Y is the correct label and G is a protected attribute.

Demographic parity

(Calders et al. 2010)



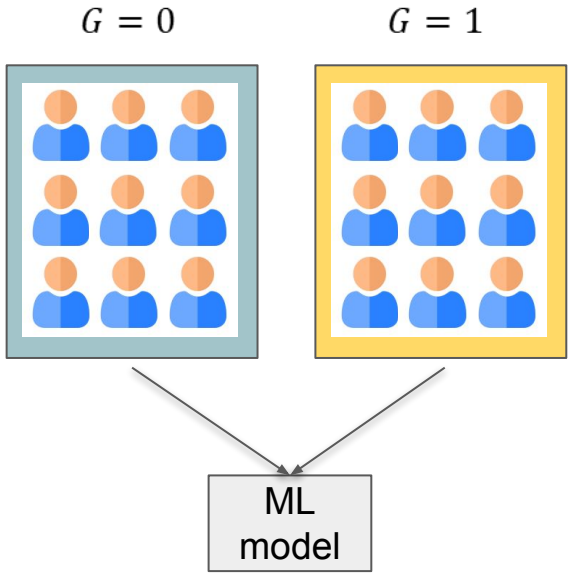
$$\mathbb{P}(\hat{Y} = 1|G = 0) = \mathbb{P}(\hat{Y} = 1|G = 1)$$

Group fairness

Classifier's decisions are *statistically independent* of the protected attribute.

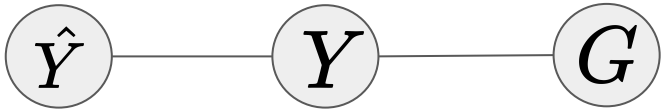
Demographic parity

(Calders et al. 2010)



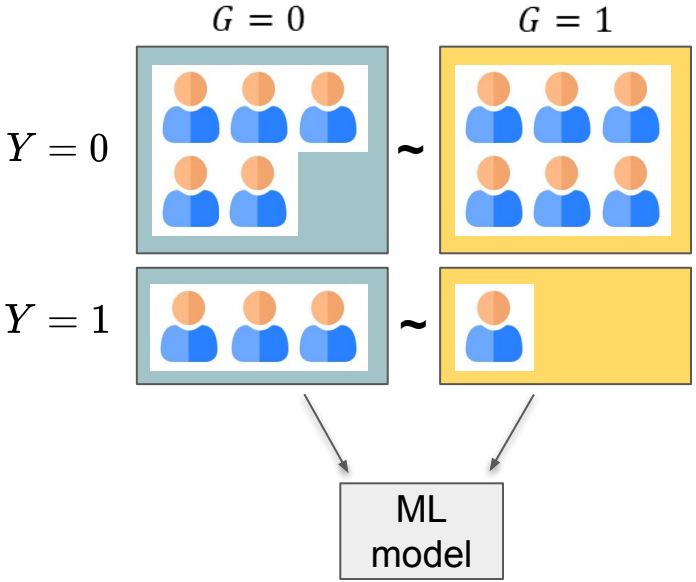
$$\mathbb{P}(\hat{Y} = 1|G = 0) = \mathbb{P}(\hat{Y} = 1|G = 1)$$

Group fairness



Classifier's decisions can only depend on protected attribute via the true label.

Equalized odds
(Hardt et al. 2016)

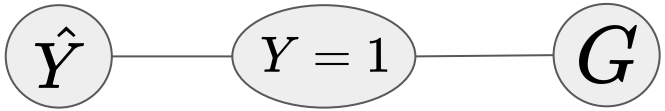


$$\mathbb{P}(\hat{Y} = 1|Y = 0, G = 0) = \mathbb{P}(\hat{Y} = 1|Y = 0, G = 1)$$

and

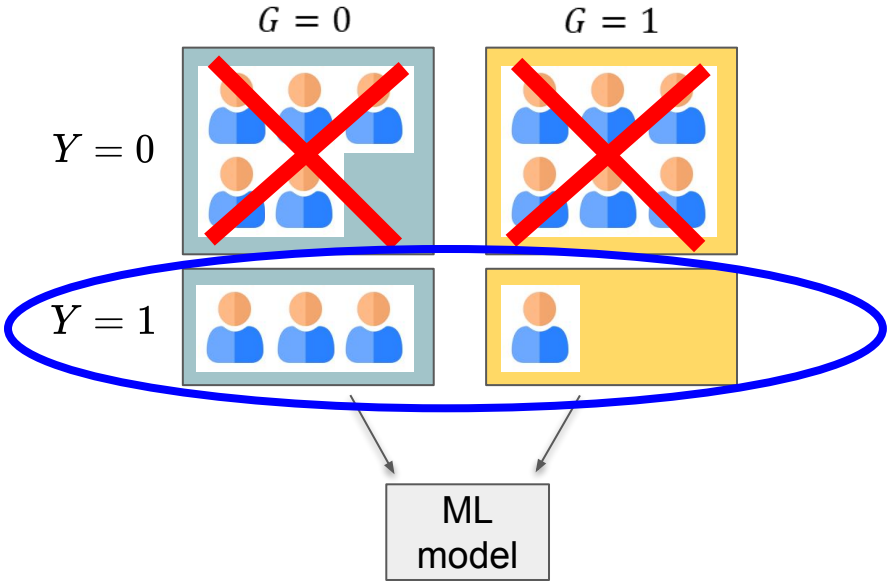
$$\mathbb{P}(\hat{Y} = 1|Y = 1, G = 0) = \mathbb{P}(\hat{Y} = 1|Y = 1, G = 1)$$

Group fairness



Restricting to positive true labels (the “advantageous” outcome), the classifier’s decisions are independent of the protected attribute.

Equality of opportunity
(Hardt et al. 2016)



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

Group fairness

Group fairness definitions

Let $h : \mathcal{X} \to \mathcal{Y}$ be a classifier and let $\mathcal{D}$ be the joint data distribution over triplets (X, G, Y) of inputs, protected attributes and labels. Then h satisfying:

(a) Demographic parity means that $h(X) \perp\!\!\!\perp G$

(b) Equalized odds means that $h(X) \perp\!\!\!\perp G \mid Y$

(c) Equality of opportunity means that $h(X) \perp\!\!\!\perp G \mid Y = 1$

Notes:

- Many other group fairness notions exist, see: Mehrabi et al., 2019
- It is possible two group fairness notions cannot hold at the same time

Counterfactual fairness (Kusner et al., 2017) - for your information, not examinable

Let $h : \mathcal{X} \to \mathcal{Y}$ be a classifier and let $\mathcal{D}$ be the data distribution generating triplets (X, G, Y) of inputs, protected attributes and labels. Then h is **counterfactually fair** if for any input x and protected attribute g:

$$\mathbb{P}(\hat{Y}_{G \leftarrow g} = y | X = x, G = g) = \mathbb{P}(\hat{Y}_{G \leftarrow g'} = y | X = x, G = g)$$

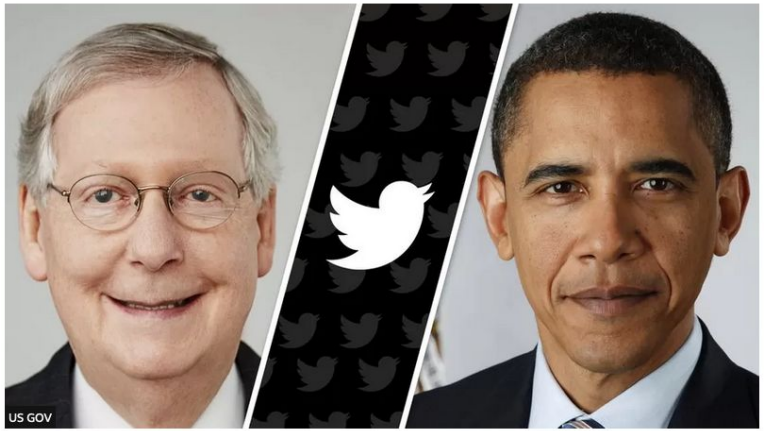
Interpretation: making an *intervention on the protected attribute* will not change the distribution of the outcome.

Group fairness asks for lack of correlation, counterfactual fairness - for lack of causation.

Fairness and bias beyond classification

Twitter investigates racial bias in image previews

21 September 2020



One user found that Twitter seemed to favour showing Mitch McConnell's face over Barack Obama's

Google search algorithms are not impartial. They can be biased, just like their designers.

Search patterns matter because sites like Google are becoming increasingly powerful arbiters of public information.

Feb. 21, 2018

YouTube's recommender AI still a horror show, finds major crowdsourced study

Natasha Lomas @riptari / 10:00 AM GMT+2 • July 7, 2021

Comment

Facebook Algorithm Shows Gender Bias in Job Ads, Study Finds

Researchers found the platform's algorithms promoted roles to certain users; company pledges to continue work in removing bias from recommendations

By Jeff Horwitz April 9, 2021