

Fairness in ML

CSE 416 Spring 2022

Sahil Verma 27 April 22

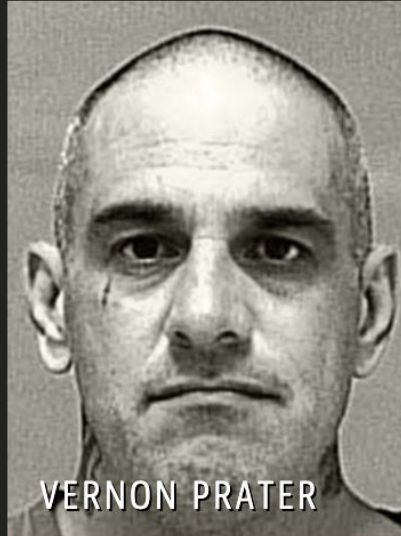
Outline

- Motivation: Examples of bias in ML in real world
- Source of Bias: Case studies
- Pillars of Trustworthy ML
- Tackling bias in ML
- Real world examples of tackling ML bias

COMPAS for Recidivism!

Propublica investigated the bias COMPAS exhibited against black people.

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



VERNON PRATER

LOW RISK

3



BRISHA BORDEN

HIGH RISK

8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Amazon algorithm for recruiting!

Amazon discovered that the algorithm for rating candidates for developer jobs was biased against women.

<https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>



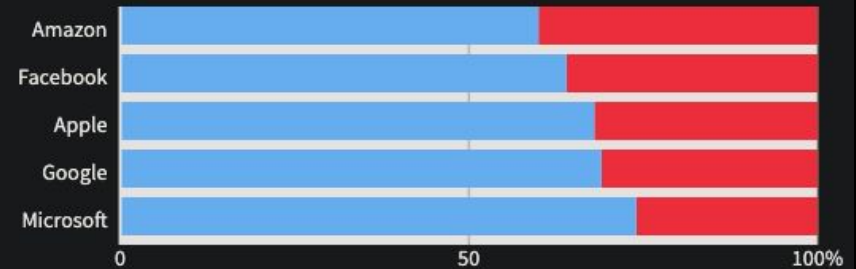
Amazon algorithm for recruiting!

Amazon discovered that the algorithm for rating candidates for developer jobs was biased against women.

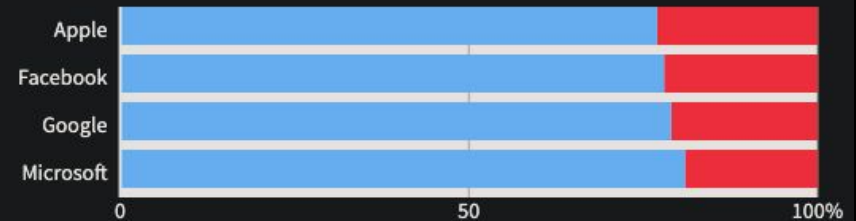
<https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>

GLOBAL HEADCOUNT

Male Female



EMPLOYEES IN TECHNICAL ROLES



Health care risk management!

The algorithm used to allocate high risk health care was found to be biased against black people. It has been used for over 200 M patients.

<https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/>



Predictive Policing!

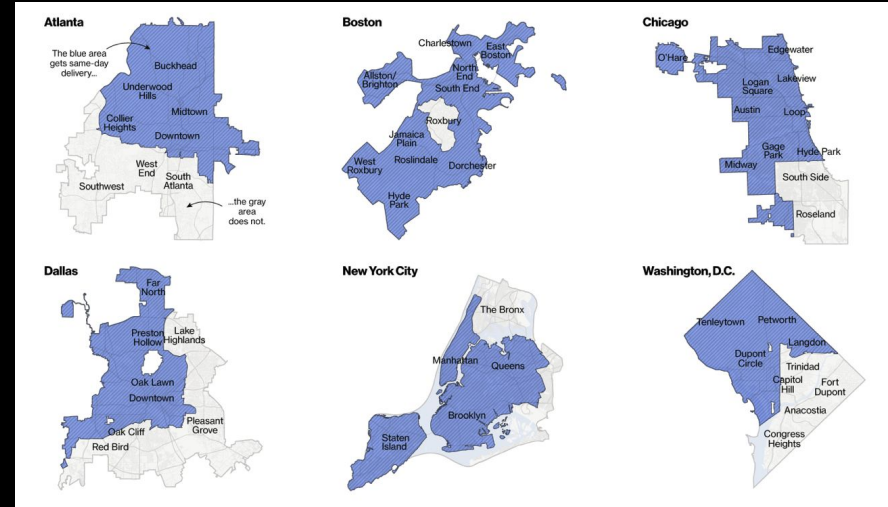
The algorithm used to predict regions with high probability of crime was found to be biased against black people.

<https://www.theverge.com/2021/12/6/22814409/go-read-this-gizmodo-analysis-predpol-software-disproportionate-algorithm>



Amazon Prime

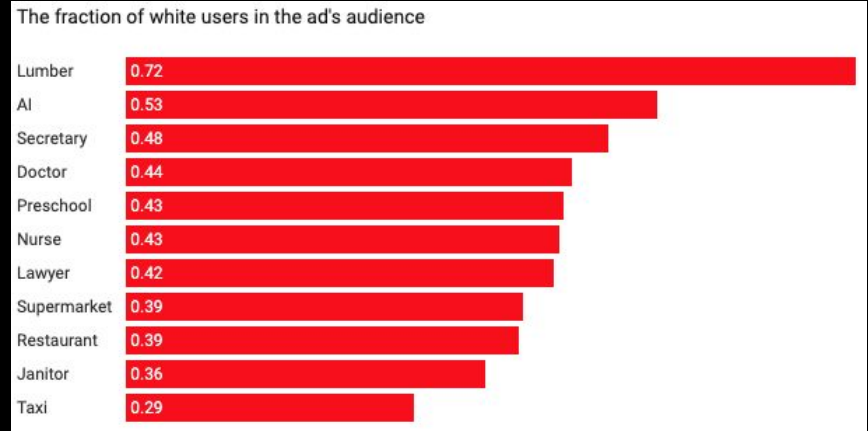
- Amazon one day delivery service was found to be biased against black neighborhoods across US.
- Most stark examples are Roxbury in Boston, Bronx in NY, Congress Heights in DC.
- Amazon used the ZIP code to decide one day delivery, not the demographics.



<https://www.bloomberg.com/graphics/2016-amazon-same-day/>

Facebook Ads

- Facebook algorithm for ads was found to be biased against women and minorities.
- Job ads for janitors and taxi drivers shown to minorities.
- Ads for nurses and secretaries shown more to women.
- Home sale jobs shown to white users, while rentals shown to minorities.



<https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias/>

Google Ads

- Latanya Sweeney, a Harvard professor found out that Google Ads were biased against blacks.
- Searching black sounding names were 25% more likely to trigger ads for criminal records than names that sound white.

Ad related to latanya sweeney ⓘ

[Latanya Sweeney Truth](#)
www.instantcheckmate.com/

Looking for **Latanya Sweeney**? Check **Latanya Sweeney's** Arrests.


Ads by Google

[Latanya Sweeney Arrested?](#)
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

[Latanya Sweeney](#)
Public Records Found For: Latanya Sweeney. View Now.
www.publicrecords.com/

[La Tanya](#)
Search for La Tanya Look Up Fast Results now!
www.ask.com/La+Tanya

(c)



iCheckmate

Latanya Sweeney
1425 Central Ave
Pittsburgh, PA 15206
DOB: [REDACTED] Sex: F Height: 5'00" Weight: 120 lbs

Criminal History See This Content ⭐⭐⭐⭐⭐
This section contains possible criminal, arrest, and criminal records for the subject of this report. When the database uses various sources or portions of arrest records, different sources have different rules regarding what information they will and will not release.

Possible Matching Arrest Records

Name	County and State	Offense	View Details
No matching arrest records were found.			

(d)

<https://www.bostonglobe.com/business/2013/02/06/harvard-professor-spots-web-search-bias/PtOgSh1ivTZMfyEGj00X4I/story.html>

Face Detection

Facial detection systems were found to have largely different accuracies for different demographic groups.





Face Detection

Every time they took a picture of each other smiling, a message flashed across the screen asking “Did someone blink?”

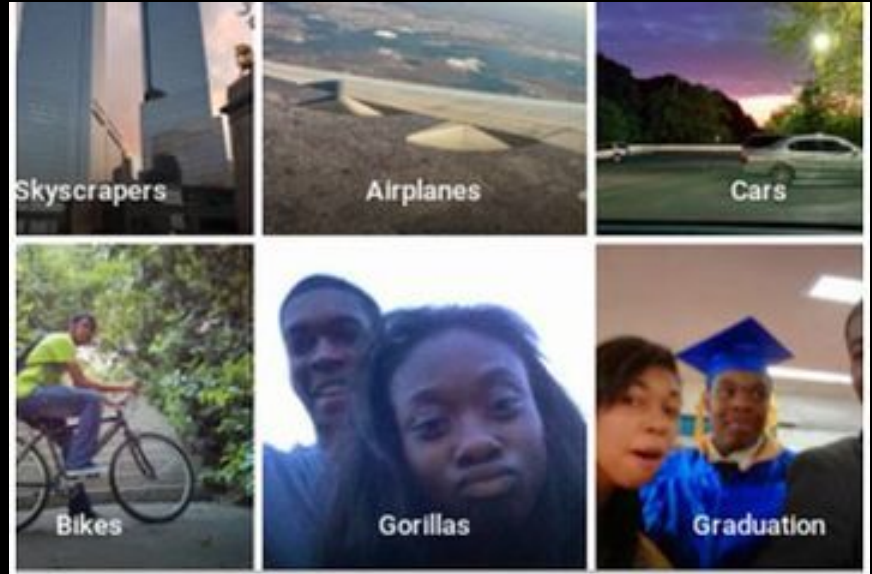
They did not blink, they are Asian.



Google Photos!

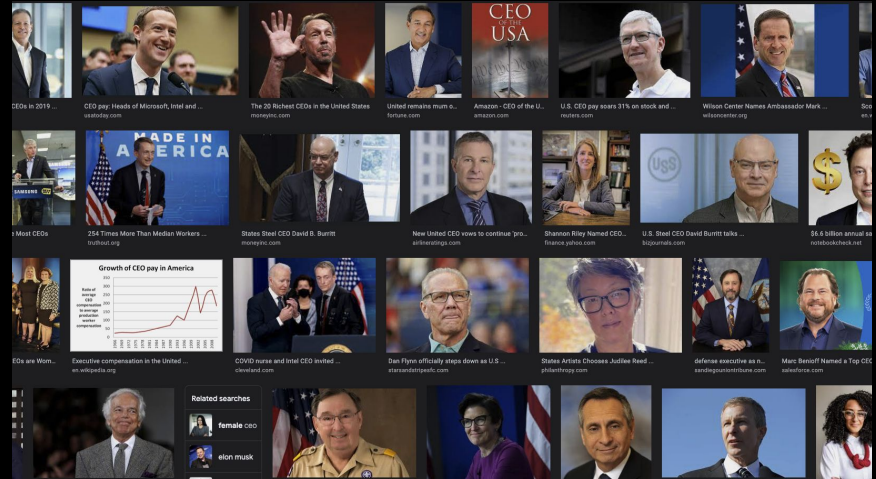
- It mistakenly tagged Black people as gorillas.
- Google fixed this problem by banning words like “Gorilla”, “Chimpanzee”, “monkeys” from Google photos.

<https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>



Google Image Search

- Image search for specific jobs yielded results that were highly skewed against minorities.
- Searching for CEO on Google found about 11% women in top 100 results instead of 27%.



<https://www.fastcompany.com/3045295/the-hidden-gender-bias-in-google-image-search>

Word Embeddings

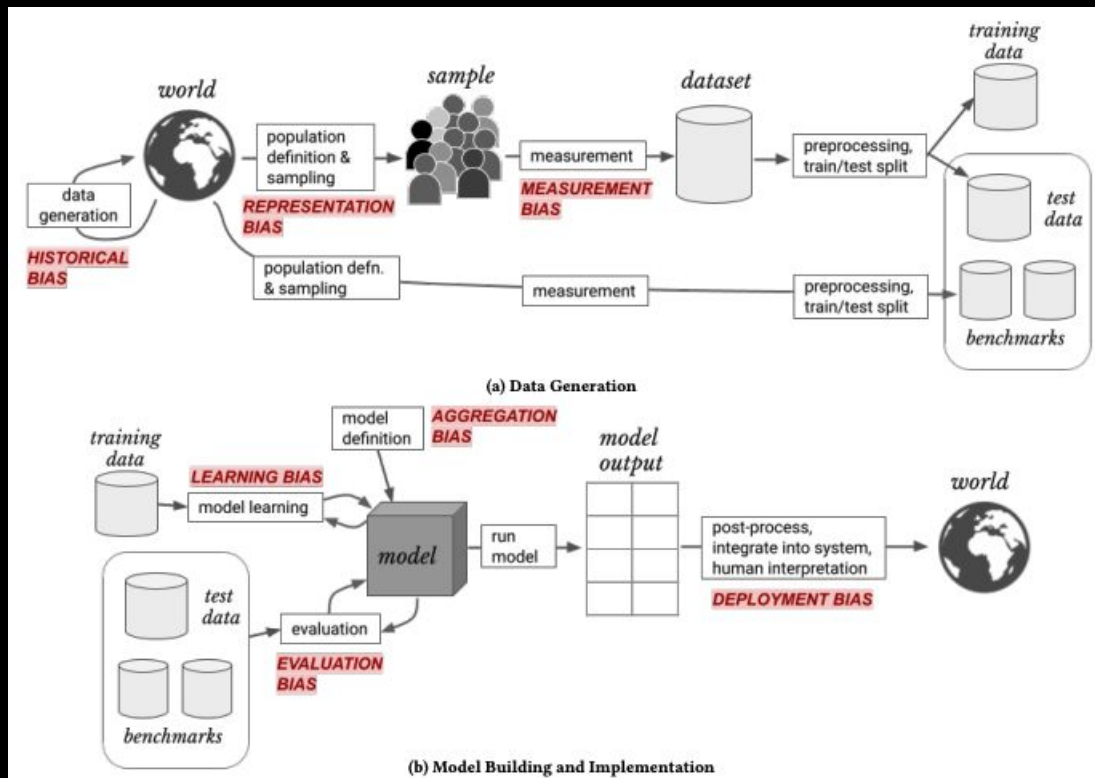
- Word embeddings used to train NLP applications like chatbots and search engines was found to be sexist.
- $\text{woman} + \text{king} - \text{man} = \text{queen}$
- $\text{woman} + \text{doctor} - \text{man} = \text{nurse}$
- $\text{programmer} + \text{woman} - \text{man} = \text{homemaker}$



Outline

- Examples of bias in ML in real world
- Source of Bias: Case studies
- Pillars of Trustworthy ML
- How to tackle bias in ML
- Real world examples of tackling ML bias

Sources of Bias: Case Studies



Case Study 1: Student Support Programs

Goal: Improve graduation rates for students

Data: Student records from different school districts and states, National Student Clearinghouse data (which gives us information about college outcomes)

Analysis: Predict risk of not graduating on time

Actions: Assign after-school programs to most at-risk students

Constraints: Resources are available to target additional tutoring to 10% of students

Case Study 2: Loans

Goal: Provide loans while balancing repayment rates for bank loans

Data: Historical loans and payments, credit reporting data, background checks

Analysis: Build model to predict risk of not repaying on time

Actions: Deny loan or increase interest rate/penalties

Case Study 3: Disaster Relief

Goal: Accurately assess damage and send appropriate relief resources

Data: Twitter posts, facebook posts geocoded with lat-long within disaster area and keywords/hashtags related to the storm

Analysis: Intensity and type of damage by neighborhood

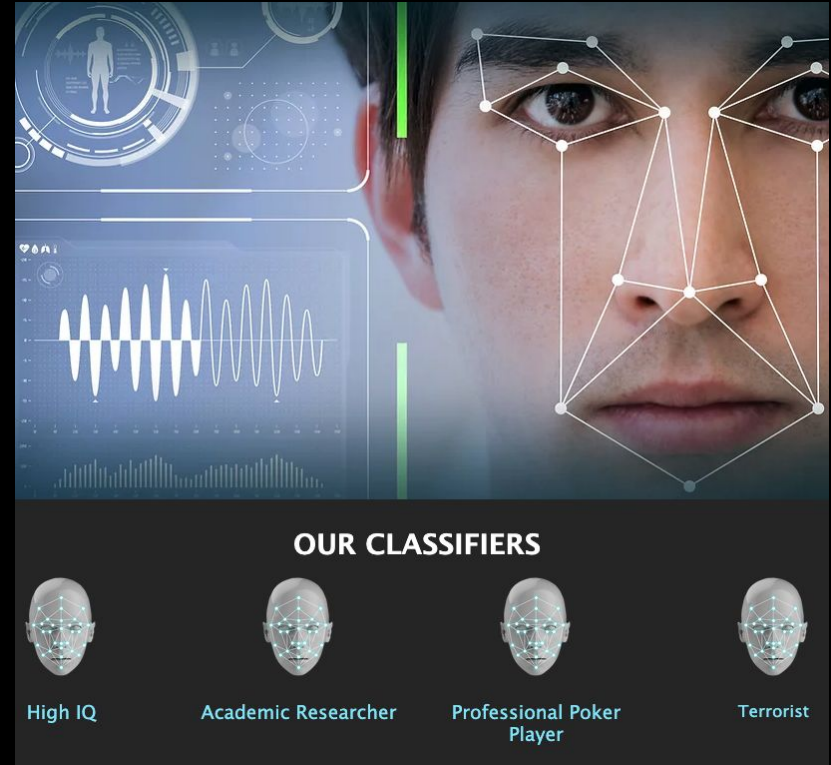
Actions: Assessment and allocate relief effort (type and amount)

Constraints: Limited resources for relief efforts

Case Study 4: Personality Detection

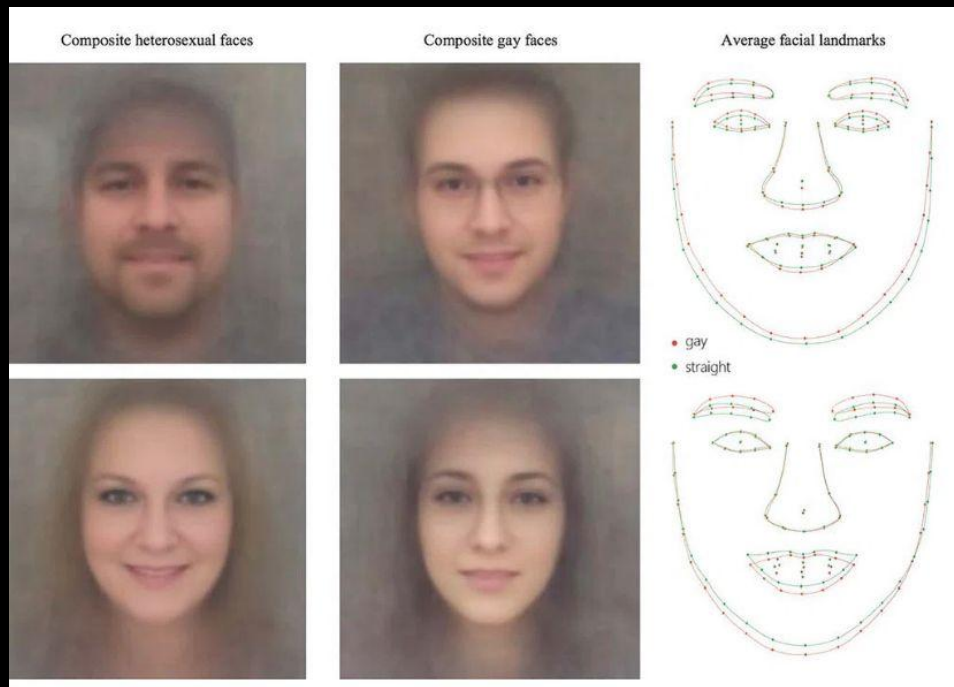
- Facepion, an Israeli startup is selling a software that can classify personalities with 80% accuracy.
- US homeland security body purchased the software in 2016.

<https://www.facepion.com/>



Case Study 5: Sexuality Detection

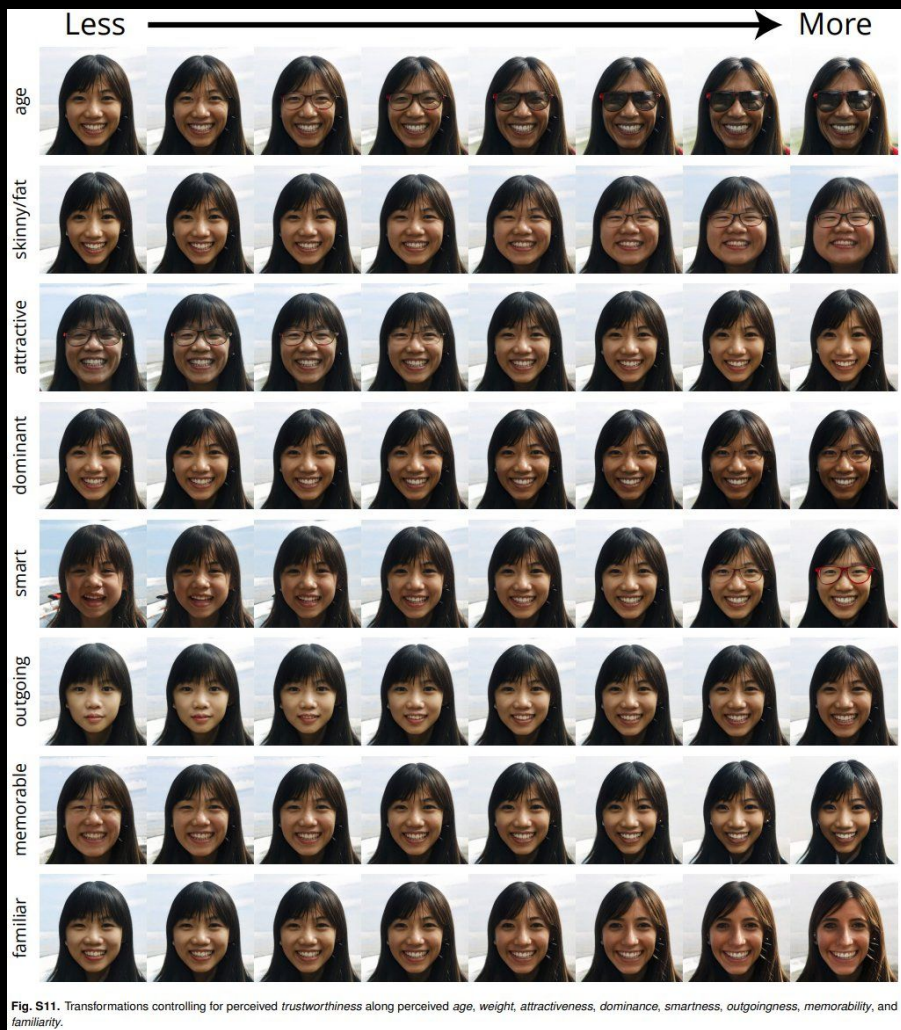
- A Stanford professor developed this facial recognition system.
- 75000 photos from online dating profiles of all white people.
- 71% accuracy for women and 81% for men. Claimed better than humans.



<https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph>

Paper published 4
days ago by
Berkeley.

[https://twitter.com/joshuacpeterson/
status/1517224918735147008](https://twitter.com/joshuacpeterson/status/1517224918735147008)



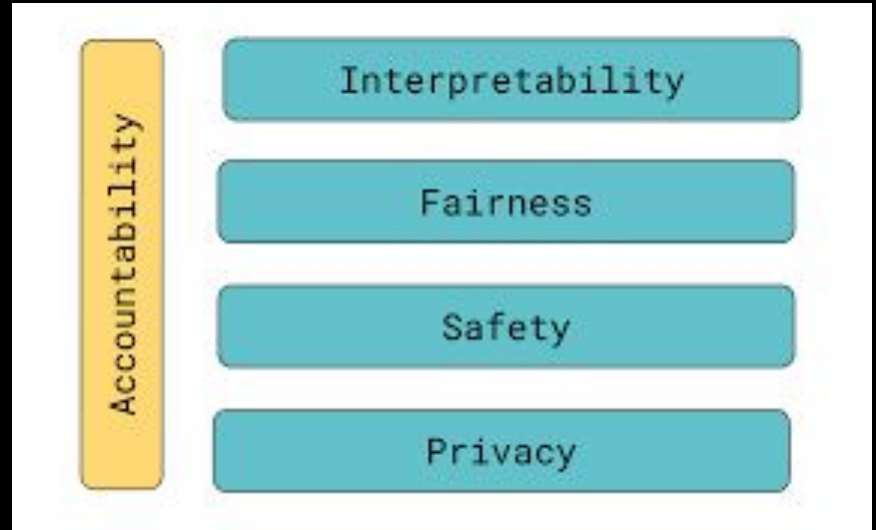
Outline

- Examples of bias in ML in real world
- Source of Bias: Case studies
- Pillars of Trustworthy ML
- How to tackle bias in ML
- Real world examples of tackling ML bias

“Since the dawn of time, any tool can be used for good or ill. Even a broom can be used to sweep the floor or hit someone over the head. The more powerful the tool, the greater the benefit or damage it can cause. While the world’s digital transformation holds great promise, the world has turned technology into both powerful tools and formidable weapons.”

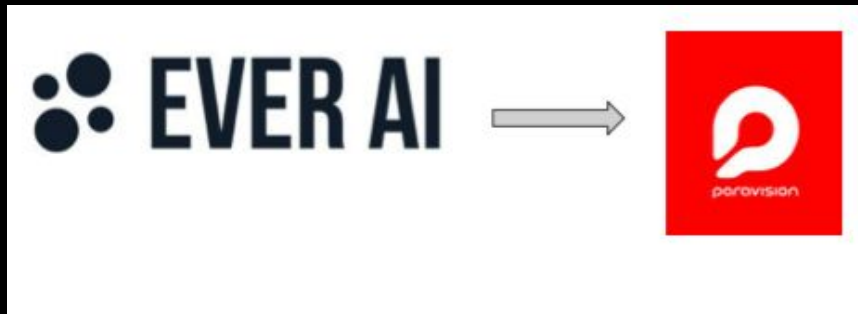
-Brad Smith
President, Microsoft

Trustworthy ML / Responsible AI



Privacy

- Cloud photo storage company.
- Trained facial recognition software using the photos people uploaded.
- Sold it to surveillance agencies.
- Renamed itself to Paravision was called out by ACLU.
- Raised \$29M before 2020.



<https://techcrunch.com/2020/08/24/ever-once-accused-of-building-facial-recognition-tech-using-customer-data-shuts-down-consumer-app/>

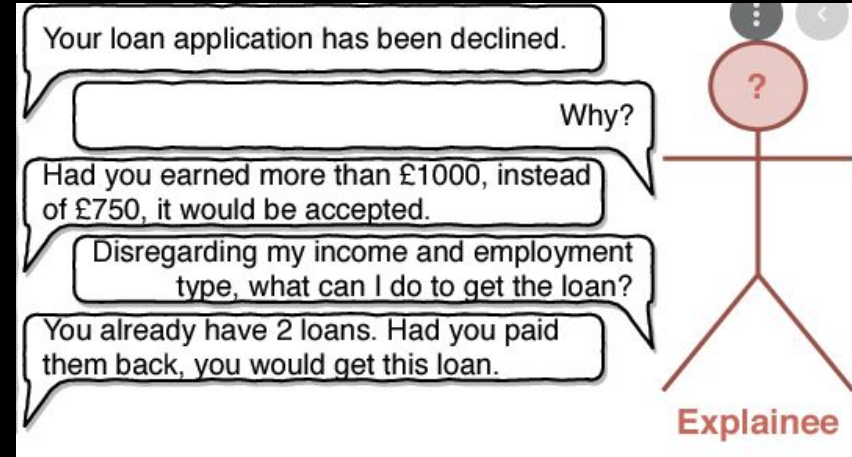
Ethics

- Builds computer vision software.
- In 2018, CEO Matt disclosed that Clarifai was selling software to Pentagon for autonomous weapons
- All due to whistleblowing by Liz
—> Parity AI



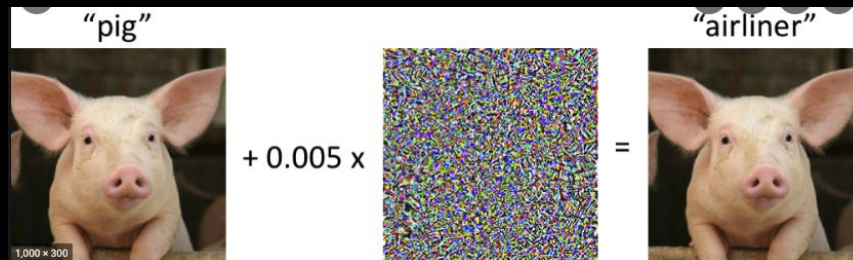
Interpretability

- A bank uses an ML model to decide which customer's loan request should be approved.
- A customer who earnestly needed money asks why did their loan request not get approved? and what can they do in order to get it approved?



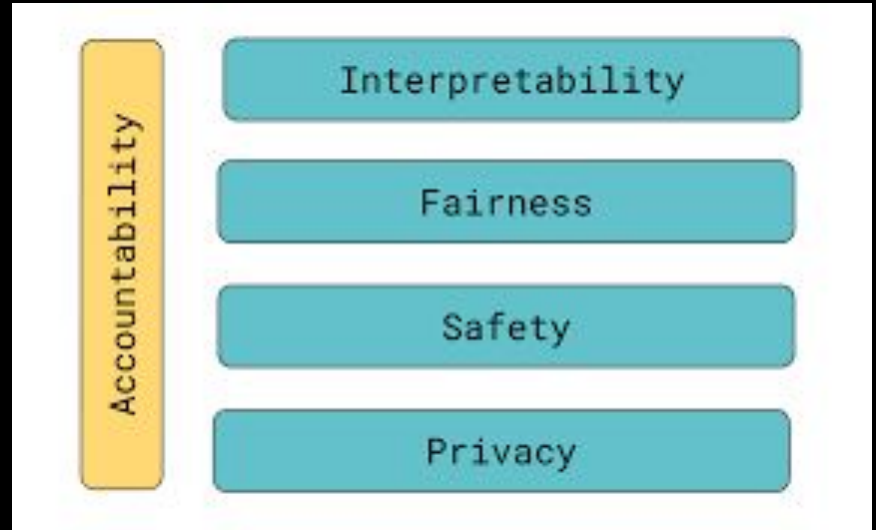
Robustness

- An image recognition model correctly recognized a pig.
- Upon addition of random noise, the ML model now thinks, it is an airplane! What do you think?



Trustworthy ML / Responsible AI

We will focus on fairness



Outline

- Examples of bias in ML in real world
- Source of Bias: Case studies
- Pillars of Trustworthy ML
- How to tackle bias in ML
- Real world examples of tackling ML bias

**How do we
develop ML
systems that
help make
decisions leading
to
fair and equitable
outcomes?**



Fairness Definitions

		True condition			
Total population		Condition positive	Condition negative	Prevalence = $\frac{\sum \text{Condition positive}}{\sum \text{Total population}}$	Accuracy (ACC) = $\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$
Predicted condition	Predicted condition positive	True positive , Power	False positive , Type I error	Positive predictive value (PPV), Precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$
	Predicted condition negative	False negative , Type II error	True negative	False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$ $F_1 \text{ score} = \frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	

Fairness Definitions Explained

Sahil Verma

Indian Institute of Technology Kanpur, India

vsahil@iitk.ac.in

Julia Rubin

University of British Columbia, Canada

mjulia@ece.ubc.ca

ABSTRACT

Algorithm fairness has started to attract the attention of researchers in AI, Software Engineering and Law communities, with more than twenty different notions of fairness proposed in the last few years. Yet, there is no clear agreement on which definition to apply in each situation. Moreover, the detailed differences between multiple definitions are difficult to grasp. To address this issue, this paper collects the most prominent definitions of fairness for the algorithmic classification problem, explains the rationale behind these definitions, and demonstrates each of them on a single unifying case-study. Our analysis intuitively explains why the same case can be considered fair according to some definitions and unfair according to others.

training data containing observations whose categories are known. We collect and clarify most prominent fairness definitions for classification used in the literature, illustrating them on a common, unifying example – the German Credit Dataset [18]. This dataset is commonly used in fairness literature. It contains information about 1000 loan applicants and includes 20 attributes describing each applicant, e.g., credit history, purpose of the loan, loan amount requested, marital status, gender, age, job, and housing status. It also contains an additional attribute that describes the classification outcome – whether an applicant has a good or a bad credit score.

When illustrating the definitions, we checked whether the classifier that uses this dataset exhibits gender-related bias. Our results were positive for some definitions and negative for others, which is

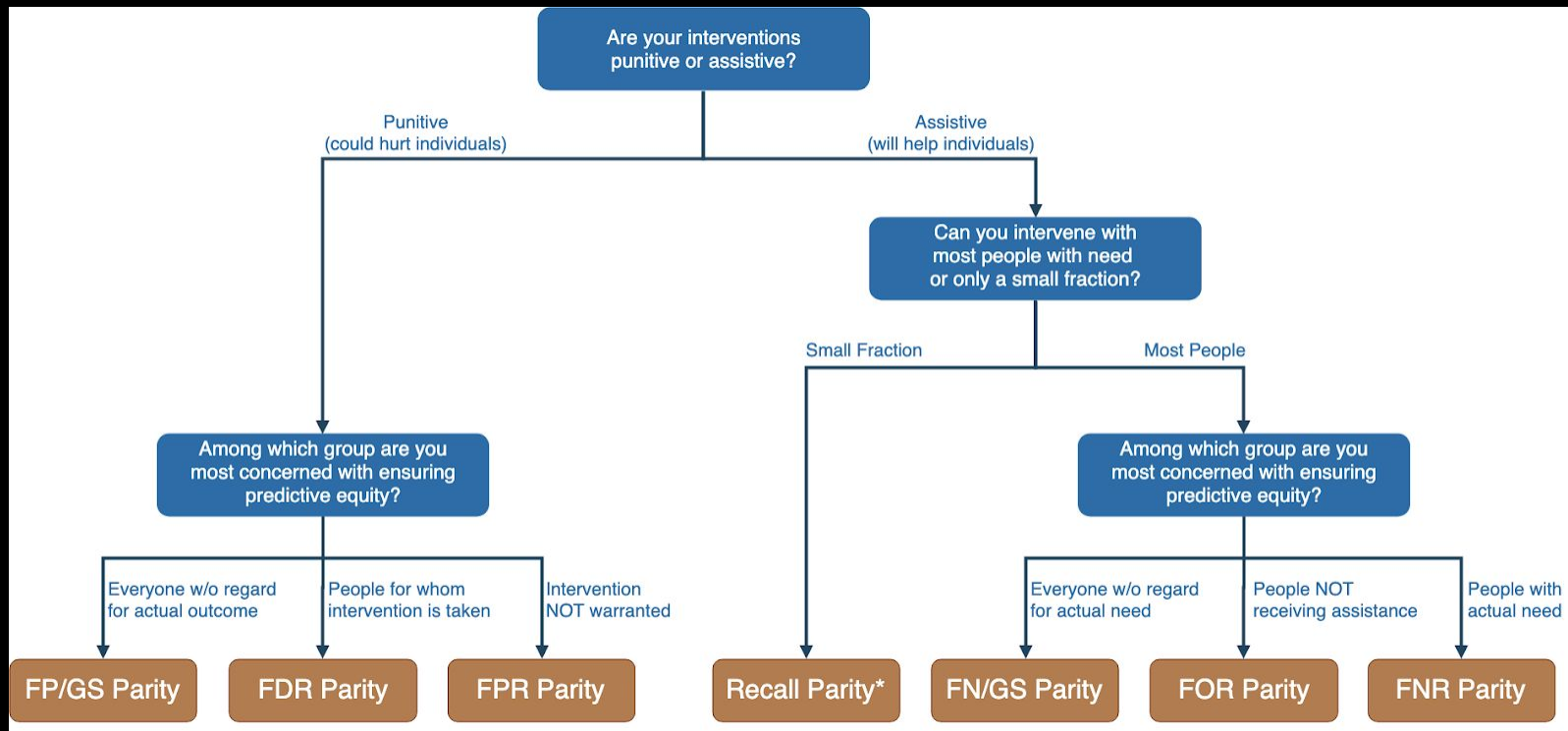
Incompatibility Between Fairness Definitions

$$FPR = \frac{p}{1 - p} \left(\frac{FDR}{1 - FDR} \right) (1 - FNR)$$

**Does that mean we
cannot achieve
fairness in ML
models?**

**In different
situations, the
appropriate
definition varies!**


Fairness Tree



Techniques for Bias Mitigation

- Pre-processing: Resampling rows of the data, reweighting rows of the data, flipping the class labels across groups, and omitting sensitive variables or proxies
- In-processing: Modify the loss function to account for fairness constraints.
- Post-processing: Adjust the outputs of the model to abide by fairness criteria.

A comparative study of fairness-enhancing interventions in machine learning

Authors:  Sorelle A. Friedler,  Carlos Scheidegger,  Suresh Venkatasubramanian,  Sonam Choudhary,
 Evan P. Hamilton,  Derek Roth [Authors Info & Claims](#)

Outline

- Examples of bias in ML in real world
- Source of Bias: Case studies
- Pillars of Trustworthy ML
- How to tackle bias in ML
- Real world examples of tackling ML bias

Facebook Ads

- Facebook gave in to the allegations and agreed to take steps:
 - Advertisers cannot target ads based on users' age, gender, race, or zip code, and other membership categories.
 - Require all advertisers to certify compliance with anti-discrimination laws.
 - Meet with ACLU members every 6 months for 3 years to enable them to monitor the reforms Facebook is undertaking.

<https://www.technologyreview.com/2019/03/20/1225/facebook-is-going-to-stop-letting-advertisers-target-by-race-gender-or-age/>

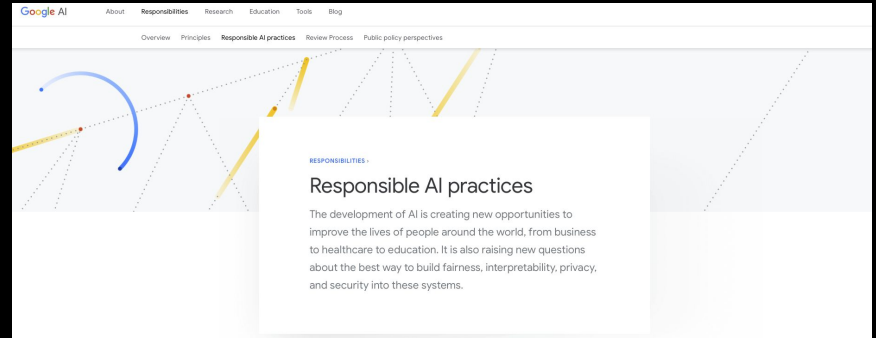
IBM abandons facial recognition

- In the aftermath of killing of George Floyd, IBM decided to abandon its facial recognition for surveillance and profiling
- IBM recognized that its software is biased and hence a hindrance in fight against racism.
- We might not have technological solutions

<https://www.bbc.com/news/technology-52978191>

Education

- Many top tech companies have made guidelines for Responsible AI practices.
- These are prescribed to the AI teams and their engineers
- These practices are (hopefully) utilized when developing their current and future products.



Inclusive Datasets

- In 2018, Google organized the inclusive dataset competition.
- They released a 500K image dataset that had datapoints across the world.

Introducing the Inclusive Images Competition

Thursday, September 6, 2018

Posted by Tulsee Doshi, Product Manager, Google AI

The release of large, publicly available image datasets, such as [ImageNet](#), [Open Images](#) and [Conceptual Captions](#), has been one of the factors driving the tremendous progress in the field of computer vision. While these datasets are a necessary and critical part of developing useful machine learning (ML) models, some open source data sets have been [found to be geographically skewed](#) based on how they were collected. Because the shape of a dataset informs what an ML model learns, such skew may cause the research community to inadvertently develop models that may perform less well on images drawn from geographical regions under-represented in those data sets. For example, the images below show one standard open-source image classifier trained on the Open Images dataset that does not properly apply “wedding” related labels to images of wedding traditions from different parts of the world.

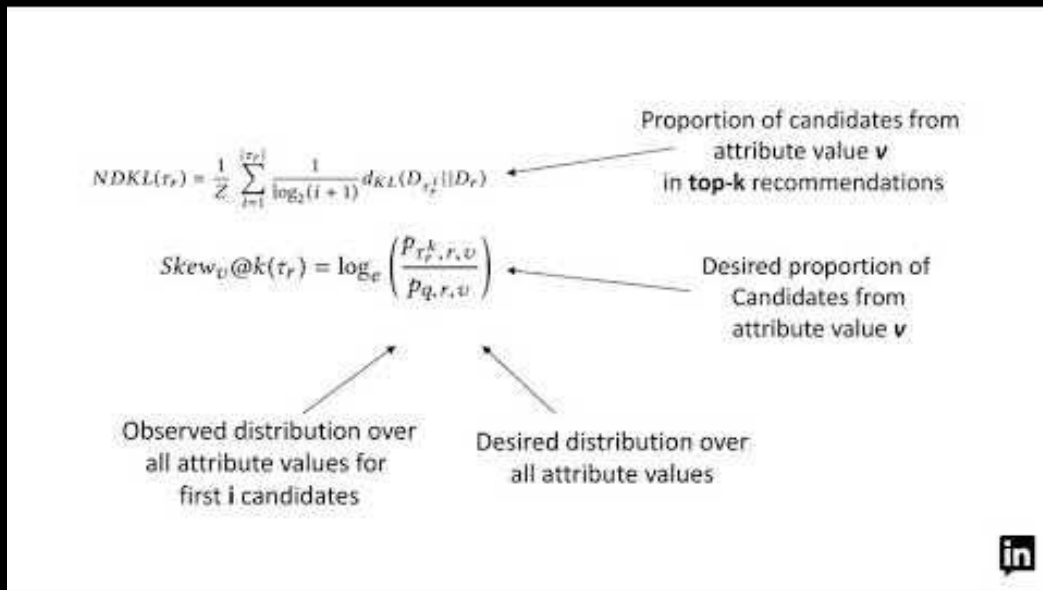


Result: Reduced Gender Bias in Translate



LinkedIn Ranking

- LinkedIn used fairness constraints to rank candidates.
- Online A/B testing resulted in three-fold increase in fairness metrics while not affecting the business metrics.



Google Dermatology

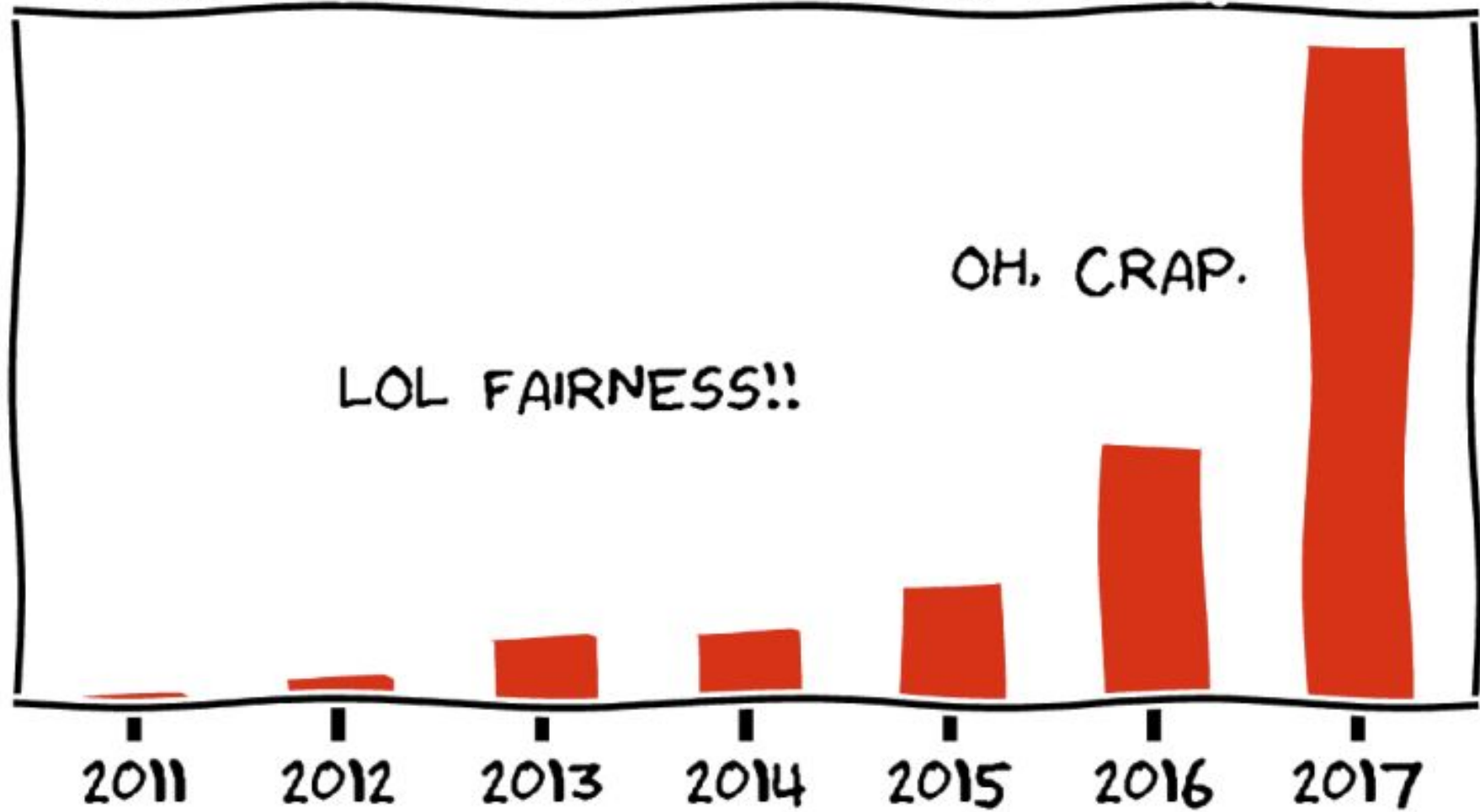
- Google released an app for detecting skin diseases in 2021
- They made every effort to collect diverse skin types.
- Even then only 3.5% of the data had skin type V and 0% of skin type VI (the dark skin types).



<https://www.forbes.com/sites/robertglatter/2021/05/21/google-announces-new-ai-app-to-diagnose-skin-conditions/>

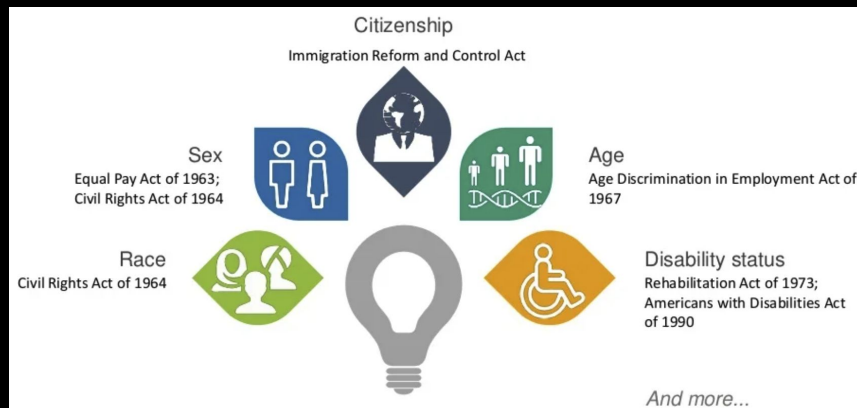
BRIEF HISTORY OF FAIRNESS IN ML

PAPERS



Regulations

- EU has been championing the law landscape of AI regulation
- GDPR enacted in 2018 requires companies to ask consent for use of personal data and several other regulations.
- US has regulations on equal credit, housing, and job opportunities. But has lot of loopholes. It's catching up!



Resources

A ton of recourses are a Google Search Away!

Courses in Fairness in ML



[Berkeley CS 294:](#)

[Fairness in machine learning](#)



[Princeton COS 597E:](#)

[Fairness in machine learning](#)



[Washington CSE 599: Foundations of](#)

[Fairness in Machine Learning](#)

- <https://adabhishekdbas.medium.com/algorithmic-bias-in-real-world-b98808e01586>
- <https://research.aimultiple.com/ai-bias/>
- <https://docs.google.com/document/d/1XnbJXELA0L3CX41MxySdPsZ-HNECxPtAw4-kZRc7OPI/edit>
- <https://twitter.com/kkenthapadi/status/1131320986303729664>
- <https://fairmlbook.org/pdf/fairmlbook.pdf>
- <https://www.kamishima.net/archive/faml.pdf>
- https://dssg.github.io/fairness_tutorial/Fairness_tutorial_slides_ic2s2_2021_version.pdf