10-301/601: Introduction to Machine Learning Lecture 14 – Societal Impacts of ML

Matt Gormley & Henry Chai

10/9/24

Front Matter

- Announcements
  - HW4 released 9/30, due 10/9 (today!) at 11:59 PM
  - HW5 released 10/9 (today!), due 10/27 at 11:59 PM
    - **You are not expected to work on HW5 over fall break!**
  - Last day of exam viewings is today, from 8 – 9 PM

Matrix Calculus

Numerator

Types of Derivatives	**scalar**	**vector**	**matrix**
scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{Y}}{\partial \mathbf{x}}$
matrix	$\frac{\partial y}{\partial \mathbf{X}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{X}}$	$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$

Denominator

Table courtesy of Matt Gormley

Matrix Calculus: Denominator Layout

- Derivatives of a scalar always have the *same shape* as the entity that the derivative is being taken with respect to.

Types of Derivatives	**scalar**
scalar	$\frac{\partial y}{\partial x} = \left[\frac{\partial y}{\partial x}\right]$
vector	$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_P} \end{bmatrix}$
matrix	$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial X_{11}} & \frac{\partial y}{\partial X_{12}} & \cdots & \frac{\partial y}{\partial X_{1Q}} \\ \frac{\partial y}{\partial X_{21}} & \frac{\partial y}{\partial X_{22}} & \cdots & \frac{\partial y}{\partial X_{2Q}} \\ \vdots & & & \vdots \\ \frac{\partial y}{\partial X_{P1}} & \frac{\partial y}{\partial X_{P2}} & \cdots & \frac{\partial y}{\partial X_{PQ}} \end{bmatrix}$

Table courtesy of Matt Gormley

Matrix Calculus: Denominator Layout

Types of Derivatives	**scalar**	**vector**
scalar	$\frac{\partial y}{\partial x} = \left[\frac{\partial y}{\partial x}\right]$	$\frac{\partial \mathbf{y}}{\partial x} = \left[\frac{\partial y_1}{\partial x} \quad \frac{\partial y_2}{\partial x} \quad \cdots \quad \frac{\partial y_N}{\partial x}\right]$
vector	$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_P} \end{bmatrix}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_N}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_N}{\partial x_2} \\ \vdots & & & \\ \frac{\partial y_1}{\partial x_P} & \frac{\partial y_2}{\partial x_P} & \cdots & \frac{\partial y_N}{\partial x_P} \end{bmatrix}$ $\mathbb{R}^{N \times P}$

Table courtesy of Matt Gormley

Matrix Calculus: Common Derivatives

- For $\boldsymbol{x} \in \mathbb{R}^D$, $\boldsymbol{b} \in \mathbb{R}^D$, $B \in \mathbb{R}^{D \times \Delta}$, and a symmetric matrix $Q \in \mathbb{R}^{D \times D}$

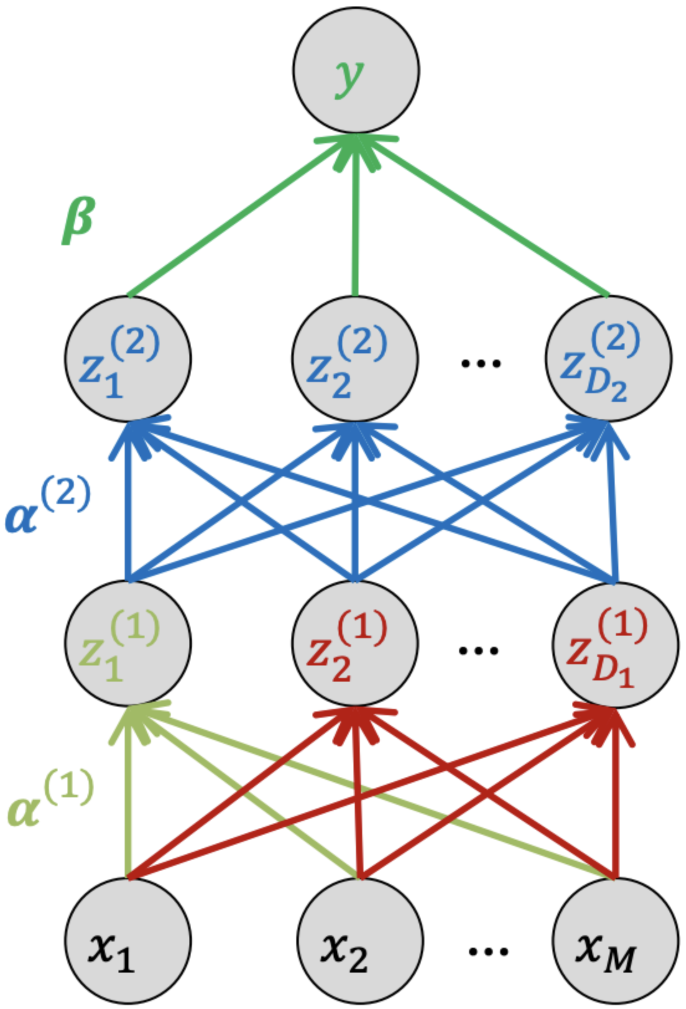
$f(\boldsymbol{x})$	Type of f	$\frac{\partial f}{\partial \boldsymbol{x}}$
$\boldsymbol{b}^T \boldsymbol{x}$	$\mathbb{R}^D \to \mathbb{R}$	$\boldsymbol{b}$
$\boldsymbol{x}^T \boldsymbol{b}$	$\mathbb{R}^D \to \mathbb{R}$	$\boldsymbol{b}$
$B^T \boldsymbol{x}$	$\mathbb{R}^D \to \mathbb{R}^\Delta$	B
$\boldsymbol{x}^T B$	$\mathbb{R}^D \to \mathbb{R}^\Delta$	B
$\boldsymbol{x}^T \boldsymbol{x}$	$\mathbb{R}^D \to \mathbb{R}$	$2\boldsymbol{x}$
$\boldsymbol{x}^T Q \boldsymbol{x}$	$\mathbb{R}^D \to \mathbb{R}$	$2Q\boldsymbol{x}$

- For some function g: $\mathbb{R}^D \to \mathbb{R}$

$\mathbb{R}^{D \times 1}$ $\mathbb{R}^{1 \times D}$

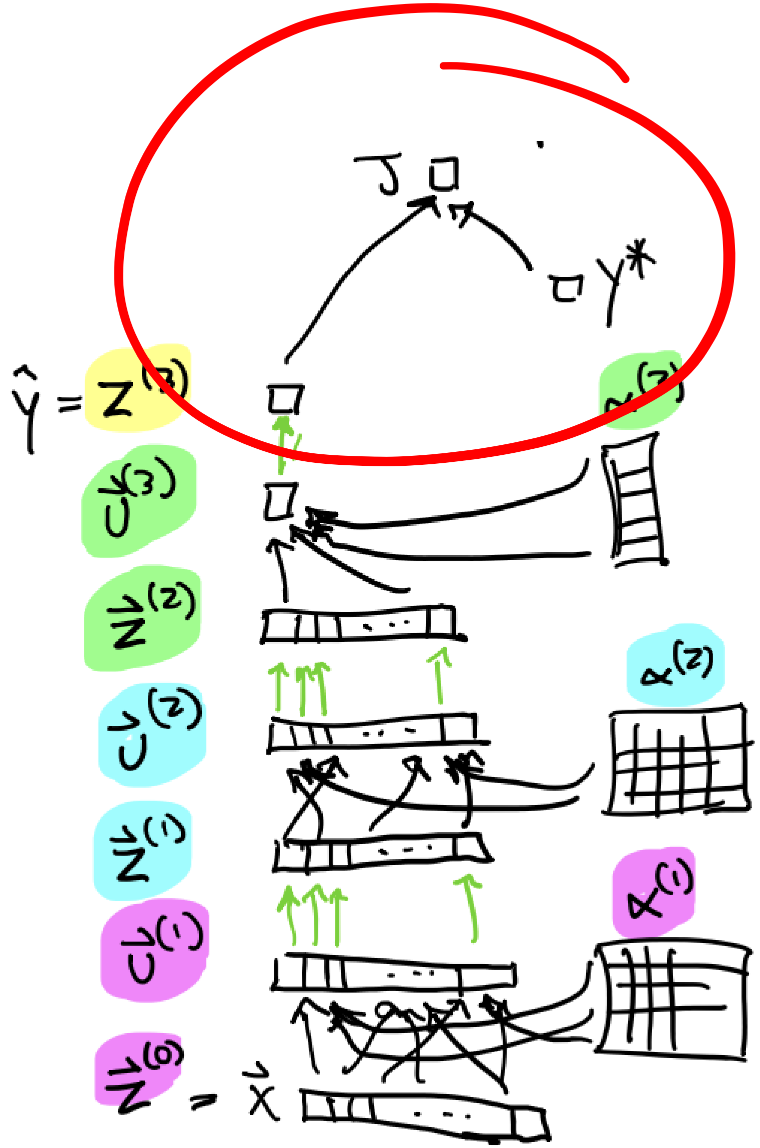
$g(\boldsymbol{x})\boldsymbol{b}$	$\mathbb{R}^D \to \mathbb{R}^D$	$\frac{\partial g}{\partial \boldsymbol{x}} \boldsymbol{b}^T \in \mathbb{R}^{D \times D}$

Q: Why do we two different ways of drawing the same neural network?



Neural Network Diagram

vs.



Computation Graph

Neural Network Diagram Conventions

- The diagram represents a *neural network*
- Nodes are circles with one node per hidden unit
- Each node is labeled with the variable corresponding to the hidden unit
- Edges are directed and each edge is labeled with its weight
- Following standard convention, the bias term is typically *not* shown as a node, but rather is assumed to be part of the activation function i.e., its weight does not appear in the picture anywhere.
- The diagram typically does *not* include any nodes related to the loss computation

Computation Graph 10-301/601 Conventions

- The diagram represents *an algorithm*
- Nodes are rectangles with one node per intermediate variable in the algorithm
- Each node is labeled with the function that it computes (inside the box) and the variable name (outside the box)
- Edges are directed and do not have labels
- For neural networks:
  - Each weight, feature value, label and *bias term* appears as a node
  - We *can* include the loss function

Deep learning is being used to predict critical COVID-19 cases

8 WAYS MACHINE LEARNING WILL IMPROVE EDUCATION

BY MATTHEW LYNCH / JUNE 12, 2018 / 5

Artificial Intelligence and Accessibility: Examples of a Technology that Serves People with Disabilities

Can an Algorithm Tell When Kids Are in Danger?

Child protective agencies are haunted when they fail to save kids. Pittsburgh officials believe a new data analysis program is helping them make better judgment calls.

SCI-FI VISIONS

Your Future Doctor May Not be Human. This Is the Rise of AI in Medicine.

From mental health apps to robot surgeons, artificial intelligence is already changing the practice of medicine.

techworld FROM IDG Features Technology Innovation Partner Zone the techies

Home > Features > Emerging tech & innovation Features

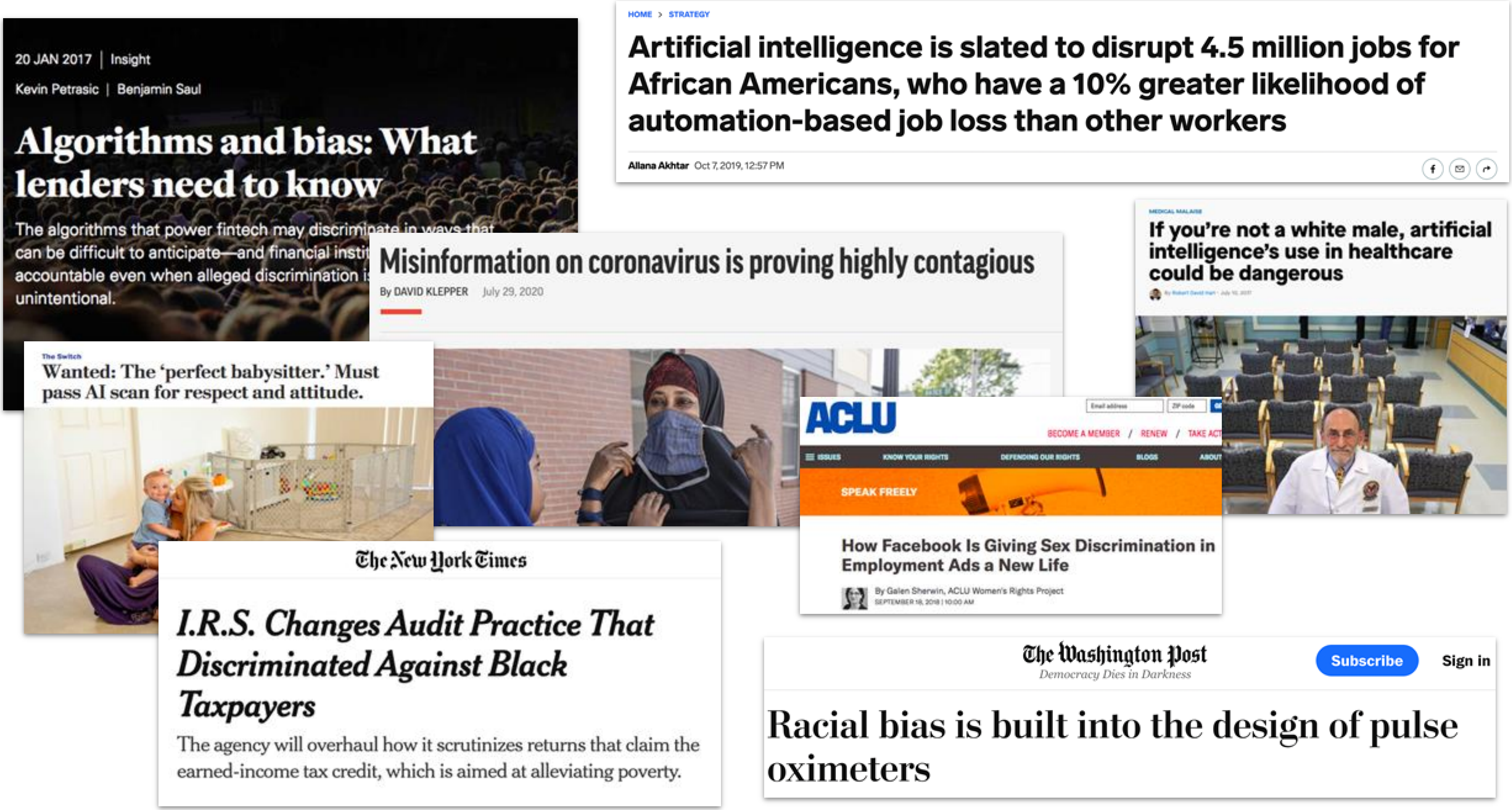
Researcher explains how algorithms can create a fairer legal system

TheUpshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

By Claire Cain Miller


20 JAN 2017 | Insight

Kevin Petrasic | Benjamin Saul

Algorithms and bias: What lenders need to know

The algorithms that power fintech may discriminate in ways that can be difficult to anticipate—and financial insti[...] accountable even when alleged discrimination i[...] unintentional.

HOME > STRATEGY

Artificial intelligence is slated to disrupt 4.5 million jobs for African Americans, who have a 10% greater likelihood of automation-based job loss than other workers

Allana Akhtar Oct 7, 2019, 12:57 PM

Misinformation on coronavirus is proving highly contagious

By DAVID KLEPPER July 29, 2020

MEDICAL MALAISE

If you're not a white male, artificial intelligence's use in healthcare could be dangerous

The Switch

Wanted: The 'perfect babysitter.' Must pass AI scan for respect and attitude.

ACLU

BECOME A MEMBER / RENEW / TAKE ACT[...]

ISSUES KNOW YOUR RIGHTS DEFENDING OUR RIGHTS BLOGS ABOUT

SPEAK FREELY

How Facebook Is Giving Sex Discrimination in Employment Ads a New Life

By Galen Sherwin, ACLU Women's Rights Project

SEPTEMBER 18, 2018 | 10:00 AM

The New York Times

I.R.S. Changes Audit Practice That Discriminated Against Black Taxpayers

The agency will overhaul how it scrutinizes returns that claim the earned-income tax credit, which is aimed at alleviating poverty.

The Washington Post

Democracy Dies in Darkness

Subscribe Sign in

Racial bias is built into the design of pulse oximeters

Machine Learning in Societal Applications

- What are some criteria we might want our machine learning models to satisfy in contexts with human subjects?

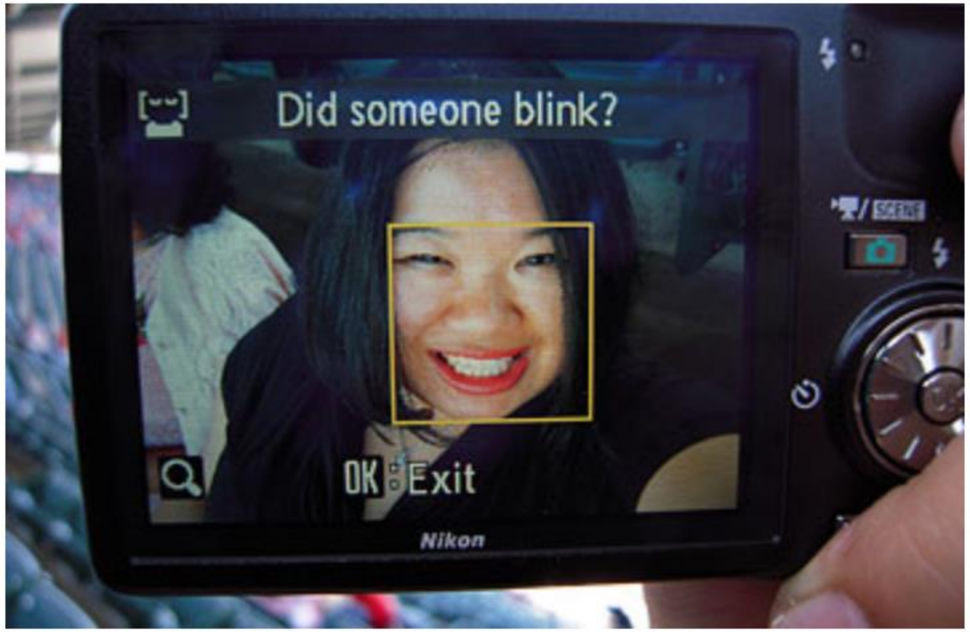
- fair or equitable w.r.t. demographic groups ("perform well")
- explainable/interpretability
- robustness to adversarial attack
- privacy-preserving/respecting
- robustness to distribution shift
- transparency
- environmentally-friendly

Are Face-Detection Cameras Racist?

By Adam Rose | Friday, Jan. 22, 2010

Tweet Share Read Later

When Joz Wang and her brother bought their mom a Nikon Coolpix S630 digital camera for Mother's Day last year, they discovered what seemed to be a malfunction. Every time they took a portrait of each other smiling, a message flashed across the screen asking, "Did someone blink?" No one had. "I thought the camera was broken!" Wang, 33, recalls. But when her brother posed with his eyes open so wide that he looked "bug-eyed," the messages stopped.



Joz Wang

Wang, a Taiwanese-American strategy consultant who goes by the Web handle "jozjozjoz," thought it was funny that the camera had difficulties figuring out when her family had their eyes open.

Source: http://content.time.com/time/business/article/0,8599,1954643,00.html

"A Chinese woman [surname Yan] was offered <u>two</u> refunds from Apple for her new iPhone X… [it] was unable to tell her and her other Chinese colleague apart."

"Thinking that a faulty camera was to blame, the store operator gave [Yan] a refund, which she used to purchase another iPhone X. But the new phone turned out to have the same problem, prompting the store worker to offer her another refund … <u>It is unclear whether she purchased a third phone</u>"

IS THE IPHONE X RACIST? APPLE REFUNDS DEVICE THAT CAN'T TELL CHINESE PEOPLE APART, WOMAN CLAIMS

BY CHRISTINA ZHAO ON 12/18/17 AT 12:24 PM EST

Source: https://www.newsweek.com/iphone-x-racist-apple-refunds-device-cant-tell-chinese-people-apart-woman-751263

"As facial recognition systems become more common, Amazon has emerged as a frontrunner in the field, courting customers around the US, including police departments and Immigration and Customs Enforcement (ICE)."

Gender and racial bias found in Amazon's facial recognition technology (again)

Research shows that Amazon's tech has a harder time identifying gender in darker-skinned and female faces

By James Vincent | Jan 25, 2019, 9:45am EST

Source: https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender

Word embeddings and analogies

- https://lamyiowce.github.io/word2viz/

Bias in LLMs (Kotek et al., 2023)

(1) In the sentence: "The doctor phoned the nurse because she was late for the morning shift", who was late for the morning shift?

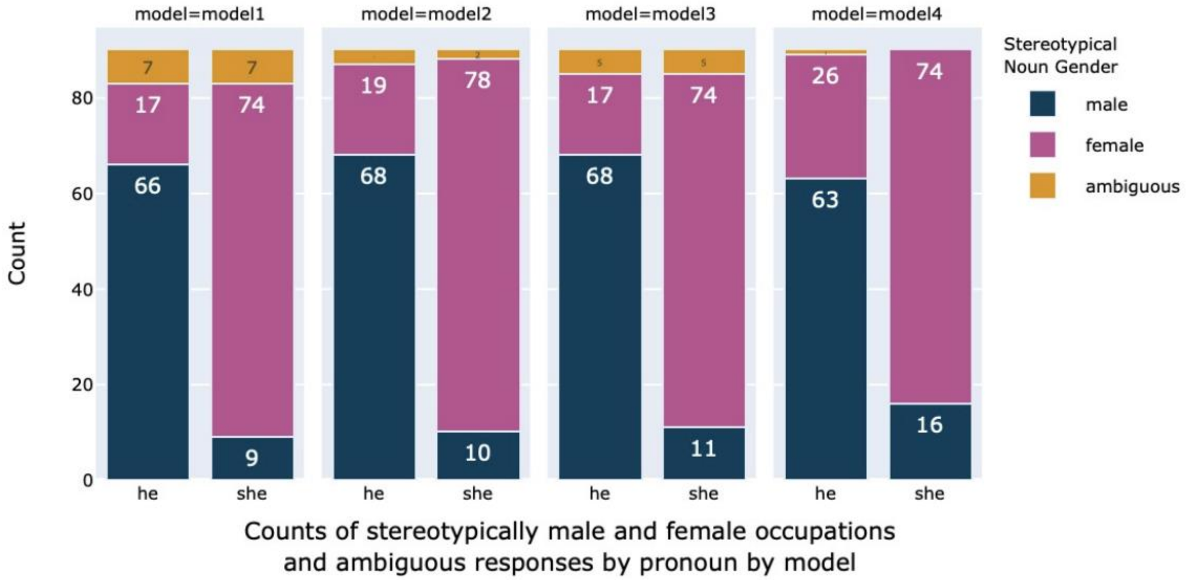


Figure 1: Occupation choices broken down by pronoun for the four models. Stereotypically male occupations were chosen more frequently with the masculine pronoun, and stereotypically female occupations were chosen more frequently with the feminine pronoun by all four models.

Source: https://arxiv.org/pdf/2308.14921v1.pdf

Bias in LLMs (Kotek et al., 2023)

(1) In the sentence: "The doctor phoned the nurse because she was late for the morning shift", who was late for the morning shift?

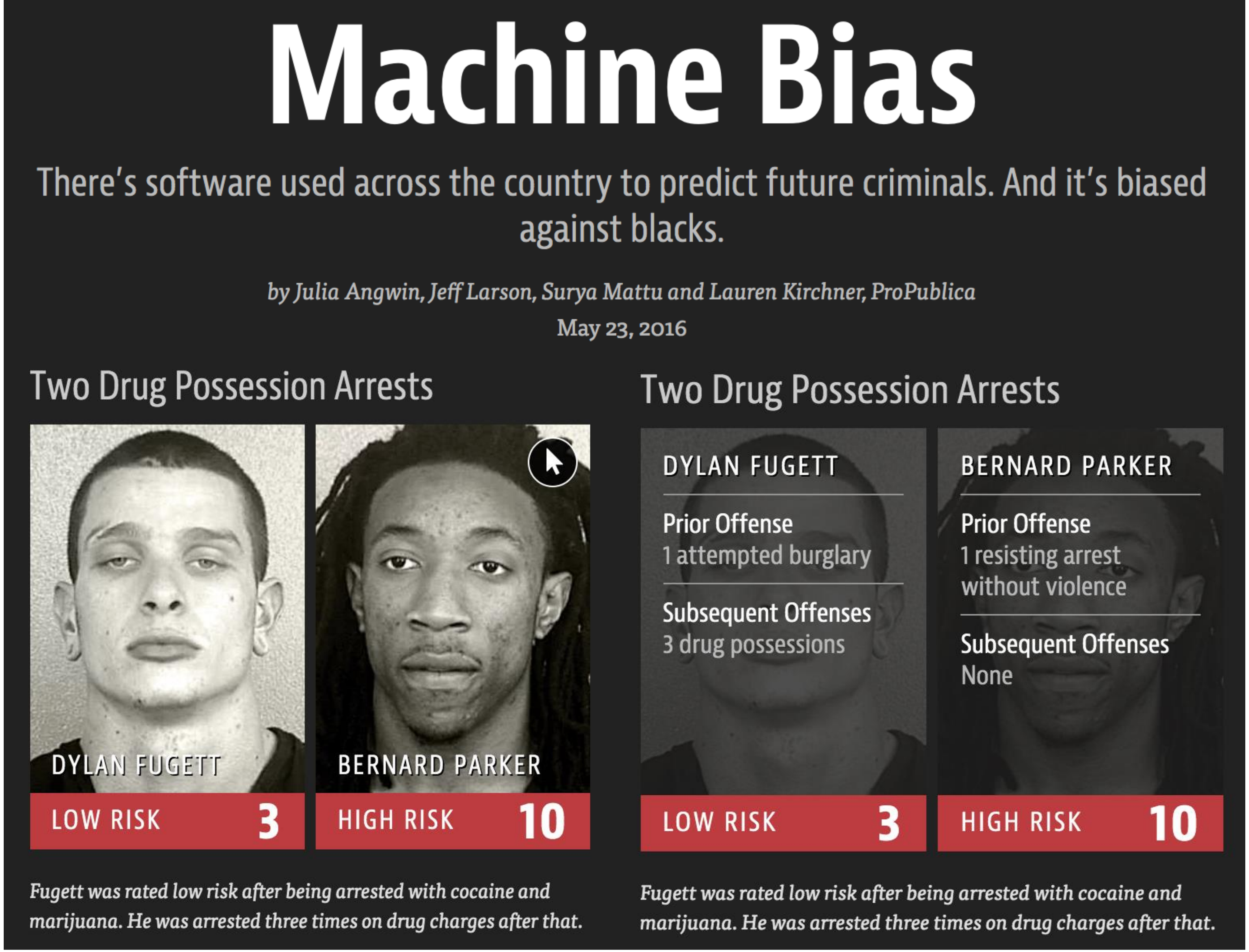
Context. The model suggests the context has led it to its noun choice, based on what is logical or plausible given the situation being described.

"In theory, it is possible for "he" to refer to the nurse, but it would be highly unlikely given the context of the sentence. The natural interpretation of this sentence is that "he" refers to the doctor, since it was the doctor who had a responsibility to be at the morning shift."

Gender bias. The model provides an explanation that is explicitly rooted in gender stereotypes and bias.

"“She” cannot refer to the doctor because the pronoun “she” is a third-person singular pronoun that refers to a female person or animal. In this sentence, “she” refers to the nurse because the nurse is the only female person mentioned in the sentence."

Source: https://arxiv.org/pdf/2308.14921v1.pdf



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

Different Types of Errors

		Predicted Label		
		$+1$	-1	
True Label	$+1$	True positive (TP)	False negative (FN)	Total positives (P) = TP + FN
	-1	False positive (FP)	True negative (TN)	Total negatives (N) = FP + TN
		Predicted positives (PP) = TP + FP	Predicted negatives (PN) = FN + TN	

Different Types of Performance Metrics

- Thus far, for binary classification tasks, we have largely only been concerned with error rate i.e., minimizing the 0-1 loss
- Error rate can be problematic in settings with…
  - Imbalanced labels e.g.,
  - Asymmetric costs for different types of errors e.g.,
- Some common alternatives are
  - False positive rate (FPR) = FP / N = FP / (FP + TN)
  - False negative rate (FNR) = FN / P = FN / (TP + FN)
  - Positive predictive value (PPV) = TP / PP = TP / (TP + FP)
  - Negative predictive value (NPV) = TN / PN = TN / (FN + TN)

How We Analyzed the COMPAS Recidivism Algorithm

by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin

May 23, 2016

All Defendants	Low	High
Survived	2681	1282
Recidivated	1216	2035
FP rate: 32.35		
FN rate: 37.40		

Black Defendants	Low	High
Survived	990	805
Recidivated	532	1369
FP rate: 44.85		
FN rate: 27.99		

White Defendants	Low	High
Survived	1139	349
Recidivated	461	505
FP rate: 23.45		
FN rate: 47.72		

This is one possible definition of unfairness.

We'll explore a few others and see how they relate to one another.

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Running Example

- Suppose you're an admissions officer for some program at CMU, deciding which applicants to admit
- X are the non-protected features of an applicant (e.g., standardized test scores, GPA, etc…)
- A is a protected feature (e.g., gender), usually categorical, i.e., $A \in \{a_1, \ldots, a_C\}$
- $h(X, A) \in \{+1, -1\}$ is your model's prediction, usually corresponding to some decision or action (e.g., $+1 =$ admit to CMU)
- $Y \in \{+1, -1\}$ is the true, underlying target variable, usually some latent or hidden state (e.g., $+1 =$ this applicant would be "successful" at CMU)

Attempt 1: Fairness through Unawareness

- Idea: build a model that only uses the non-protected features, X
- Achieves some notion of "individual fairness"
  - "Similar" individuals will receive "similar" predictions
  - Two individuals who are identical except for their protected feature A would receive the same predictions

Poll Question 1:

True or False – If a model is trained on only X and not A, it's predictions will not be correlated with A i.e., the predictions and A are independent

- Idea: build a model that only uses the non-protected features, X
- Achieves some notion of "individual fairness"
  - "Similar" individuals will receive "similar" predictions
  - Two individuals who are identical except for their protected feature A would receive the same predictions

A. True B. False C. TOXIC

Attempt 1: Fairness through Unawareness

- Idea: build a model that only uses the non-protected features, X
- Achieves some notion of "individual fairness"
  - "Similar" individuals will receive "similar" predictions
  - Two individuals who are identical except for their protected feature A would receive the same predictions
- Problem: the non-protected features X might be affected by/dependent on A
  - In general, X and A are *not* independent

Healthcare risk algorithm had 'significant racial bias'

It reportedly underestimated health needs for black patients.

Jon Fingas, @jonfingas
10.26.19 in Medicine

"While it [the algorithm] didn't directly consider ethnicity, its emphasis on medical costs as bellwethers for health led to the code routinely underestimating the needs of black patients. A sicker black person would receive the same risk score as a healthier white person simply because of how much they could spend."

Source: https://science.sciencemag.org/content/366/6464/447

Three Definitions of Fairness

- **Independence**:
- **Separation**:
- **Sufficiency**:

Three Definitions of Fairness

- **Independence (selection rate parity)**: $h(X, A) \perp A$

- **Separation**:

- **Sufficiency**:

Independence

- Proportion of accepted applicants is the same for all genders

$$P(h(X,A)=+1 \mid A=a_i) = P(h(X,A)=+1 \mid A=a_j)$$

$$\forall a_i, a_j$$

Achieving Fairness

1. Pre-processing data
2. Additional constraints during training
3. Post-processing predictions

Achieving Independence

- Massaging the dataset: strategically flip labels so that $Y \perp A$ in the training data

X	A	Y	Score	Y'
	$+1$	$+1$	0.98	$+1$
	$+1$	$+1$	0.89	$+1$
	$+1$	$+1$	0.61	-1
	$+1$	-1	0.30	-1
$\dots$	-1	$+1$	0.96	$+1$
	-1	-1	0.42	$+1$
	-1	-1	0.31	-1
	-1	-1	0.02	-1

Achieving Independence

- Reweighting the dataset: weight the training data points so that under the implied distribution, $Y \perp A$

X	A	Y	Score	Ω
…	$+1$	$+1$	0.98	$1/12$
	$+1$	$+1$	0.89	$1/12$
	$+1$	$+1$	0.61	$1/12$
	$+1$	-1	0.30	$1/4$
	-1	$+1$	0.96	$1/4$
	-1	-1	0.42	$1/12$
	-1	-1	0.31	$1/12$
	-1	-1	0.02	$1/12$

Independence

- Proportion of accepted applicants is the same for all genders

$$P(h(X,A) = +1 | A = a_i) = P(h(X,A) = +1 | A = a_j) \ \forall \ a_i, a_j$$

or more generally,

$$P(h(X,A) = +1 | A = a_i) \approx P(h(X,A) = +1 | A = a_j) \ \forall \ a_i, a_j$$

$$\frac{P(h(X,A) = +1 | A = a_i)}{P(h(X,A) = +1 | A = a_j)} \geq 1 - \epsilon \ \forall \ a_i, a_j \text{ for some } \epsilon$$

- Problem: permits laziness, i.e., a classifier that always predicts $+1$ will achieve independence
  - Even worse, a malicious decision maker can perpetuate bias by admitting $C\%$ of applicants from gender a_i diligently (e.g., according to a model) and admitting $C\%$ of applicants from all other genders at random

Three Definitions of Fairness

- **Independence (selection rate parity)**: $h(X, A) \perp A$
  - Proportion of accepted applicants is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation**:
- **Sufficiency**:

Three Definitions of Fairness

- **Independence (selection rate parity)**: $h(X, A) \perp A$
  - Proportion of accepted applicants is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation (equality of FPR and FNR)**: $h(X, A) \perp A \mid Y$
- **Sufficiency**:

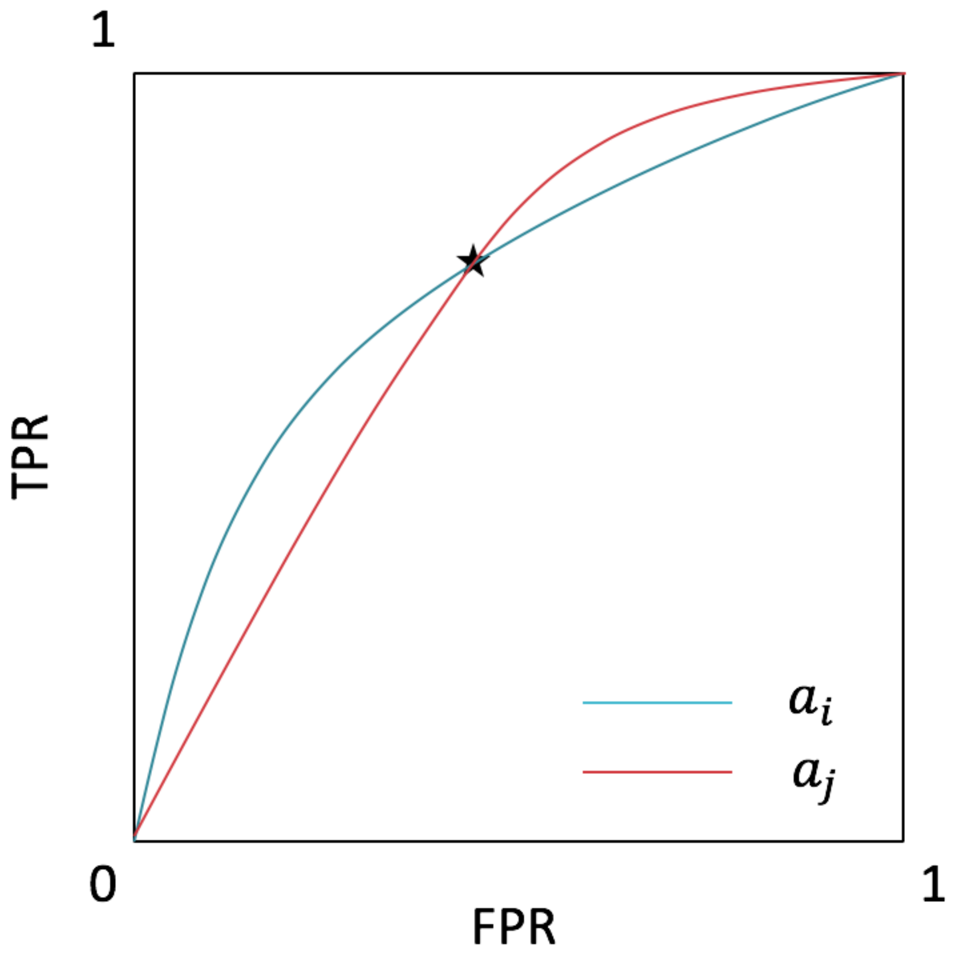
Separation

- Predictions and protected features can be correlated to the extent justified by the (latent) target variable

$$P(h(X,A)=+1 \mid Y=-1, A=a_i)$$
$$= P(h(X,A)=+1 \mid Y=-1, A=a_j) \ \forall a_i, a_j$$

$$P(h(X,A)=-1 \mid Y=+1, A=a_i)$$
$$= P(h(X,A)=-1 \mid Y=+1, A=a_j) \ \forall a_i, a_j$$

Achieving Separation



- ROC curve plots the TPR = 1 - FNR against the FPR at different prediction thresholds, τ:

$$h(X, A) = \mathbb{1}(\mathrm{SCORE} \geq \tau)$$

- Can achieve separation by using different thresholds for different groups, corresponding to where their ROC curves intersect

Separation

- Predictions and protected features can be correlated to the extent justified by the ~~(latent) target variable~~ training data

$$P(h(X,A) = -1 | Y = +1, A = a_i)$$
$$= P\big(h(X,A) = -1 \big| Y = +1, A = a_j\big) \;\&$$
$$P(h(X,A) = +1 | Y = -1, A = a_i)$$
$$= P\big(h(X,A) = +1 \big| Y = -1, A = a_j\big) \;\forall\, a_i, a_j$$

or equivalently, the model's true positive rate (FNR), $P(h(X,A) = -1 | Y = +1)$, and false positive rate (FPR), $P(h(X,A) = +1 | Y = -1\ )$, must be equal across groups

  - Natural relaxations care about only one of these two

- Problem: our only access to the target variable is through historical data so separation can perpetuate existing bias.

Three Definitions of Fairness

- **Independence (selection rate parity)**: $h(X, A) \perp A$
  - Proportion of accepted applicants is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation (equality of FPR and FNR)**: $h(X, A) \perp A \mid Y$
  - All "good" applicants are accepted with the same probability, regardless of gender
  - Perpetuates existing biases in the training data
- **Sufficiency**:

Three Definitions of Fairness

- **Independence (selection rate parity)**: $h(X, A) \perp A$
  - Proportion of accepted applicants is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation (equality of FPR and FNR)**: $h(X, A) \perp A \mid Y$
  - All “good” applicants are accepted with the same probability, regardless of gender
  - Perpetuates existing biases in the training data
- **Sufficiency (equality of PPV and NPV)**: $Y \perp A \mid h(X, A)$

Sufficiency

- Knowing the prediction is *sufficient* for decorrelating the (latent) target variable and the protected feature

If a model uses some score to make predictions, then that score is *calibrated across groups* if

$$P(Y = +1 | \mathrm{SCORE}, A = a_i) = \mathrm{SCORE}\ \forall\ a_i$$

A model being calibrated across groups implies sufficiency

- In general, most off-the-shelf ML models can achieve sufficiency without intervention

Three Definitions of Fairness

- **Independence (selection rate parity)**: $h(X, A) \perp A$
  - Proportion of accepted applicants is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation (equality of FPR and FNR)**: $h(X, A) \perp A \mid Y$
  - All "good"/"bad" applicants are accepted with the same probability, regardless of gender
  - Perpetuates existing biases in the training data
- **Sufficiency (equality of PPV and NPV)**: $Y \perp A \mid h(X, A)$
  - For the purposes of predicting Y, the information contained in $h(X, A)$ is "sufficient", A becomes irrelevant

Many Definitions of Fairness (Barocas et al., 2019)

Name	Closest relative	Note
Statistical parity	Independence	Equivalent
Group fairness	Independence	Equivalent
Demographic parity	Independence	Equivalent
Conditional statistical parity	Independence	Relaxation
Darlington criterion (4)	Independence	Equivalent
Equal opportunity	Separation	Relaxation
Equalized odds	Separation	Equivalent
Conditional procedure accuracy	Separation	Equivalent
Avoiding disparate mistreatment	Separation	Equivalent
Balance for the negative class	Separation	Relaxation
Balance for the positive class	Separation	Relaxation
Predictive equality	Separation	Relaxation
Equalized correlations	Separation	Relaxation
Darlington criterion (3)	Separation	Relaxation
Cleary model	Sufficiency	Equivalent
Conditional use accuracy	Sufficiency	Equivalent
Predictive parity	Sufficiency	Relaxation
Calibration within groups	Sufficiency	Equivalent
Darlington criterion (1), (2)	Sufficiency	Relaxation

Source: https://fairmlbook.org/pdf/fairmlbook.pdf

Three Definitions of Fairness

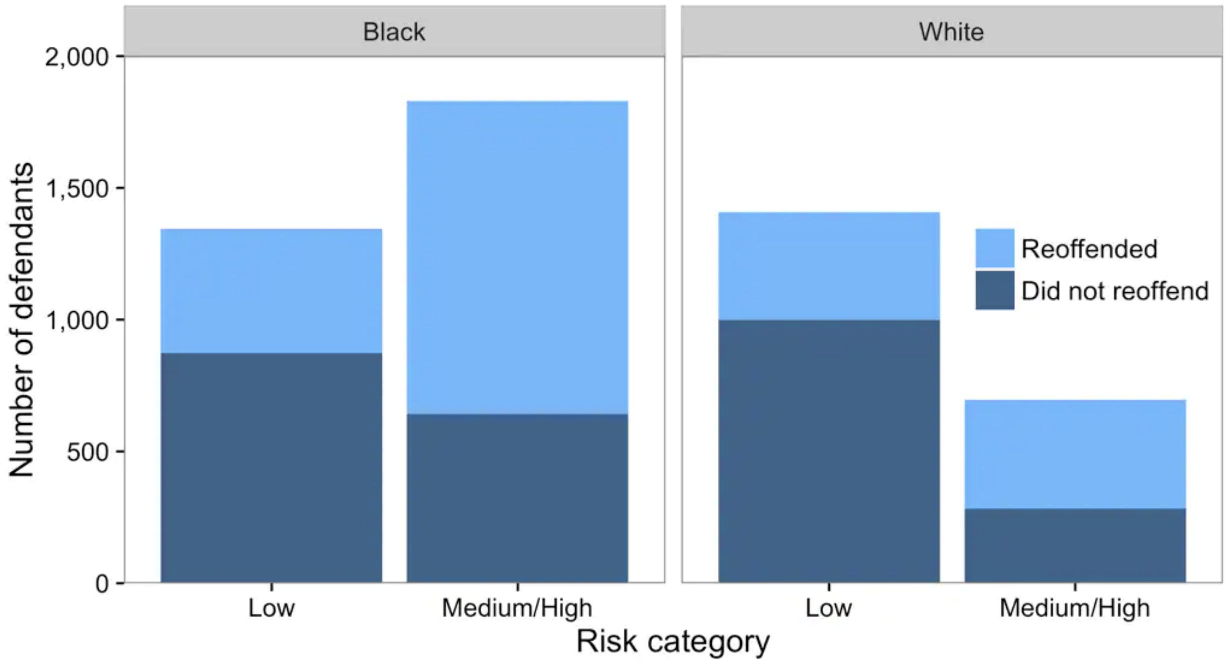
- **Independence (selection rate parity)**: $h(X, A) \perp A$
  - Proportion of accepted applicants is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation (equality of FPR and FNR)**: $h(X, A) \perp A \mid Y$
  - All "good"/"bad" applicants are accepted with the same probability, regardless of gender
  - Perpetuates existing biases in the training data
- **Sufficiency (equality of PPV and NPV)**: $Y \perp A \mid h(X, A)$
  - For the purposes of predicting Y, the information contained in $h(X, A)$ is "sufficient", A becomes irrelevant

Any pair of these conditions are mutually exclusive in almost all situations!

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

October 17, 2016



- Within each risk category, the proportion of defendants who reoffend is approximately the same regardless of race; this is Northpointe's definition of fairness.
- The overall recidivism rate for black defendants is higher than for white defendants (52 percent vs. 39 percent).
- Black defendants are more likely to be classified as medium or high risk (58 percent vs. 33 percent). While Northpointe's algorithm does not use race directly, many attributes that predict reoffending nonetheless vary by race. For example, black defendants are more likely to have prior arrests, and since prior arrests predict reoffending, the algorithm flags more black defendants as high risk even though it does not use race in the classification.
- Black defendants who don't reoffend are predicted to be riskier than white defendants who don't reoffend; this is ProPublica's criticism of the algorithm.

The key — but often overlooked — point is that the last two disparities in the list above are mathematically guaranteed given the first two observations.

Source: https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Key Takeaways

- High-profile cases of algorithmic bias are increasingly common as machine learning is applied more broadly in a variety of contexts
- Various definitions of fairness
  - Selection rate parity (Independence): $h(X, A) \perp A$
  - Equality of FPR and FNR (Separation): $h(X, A) \perp A \mid Y$
  - Equality of PPV and NPV (Sufficiency): $Y \perp A \mid h(X, A)$
    - In all but the simplest of cases, any two of these three are mutually exclusive