

10-701: Introduction to Machine Learning Lecture 19: Algorithmic Bias

Pradeep Ravikumar & Geoff Gordon

Spring 2026

Front Matter

- Link for Midsemester feedback:



- Announcements
 - Homework 3 quiz will be on Friday. There will not be a recitation this week.
 - Project proposal is due on Friday, which requires a group of 2-3 for submission.

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

8 WAYS MACHINE LEARNING WILL IMPROVE EDUCATION

BY MATTHEW LYNCH / JUN 12, 2018 / 5



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



The New York Times 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

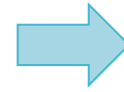
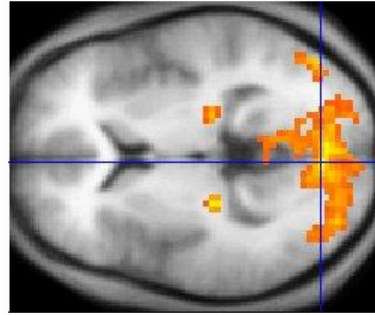
ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

By Claire Cain Miller

Machine Learning in Action

- Decoding thoughts from brain scans



Rob a bank ...

[Home](#) » [Health & Wellness](#)

Brain Scans: Are You a Criminal?

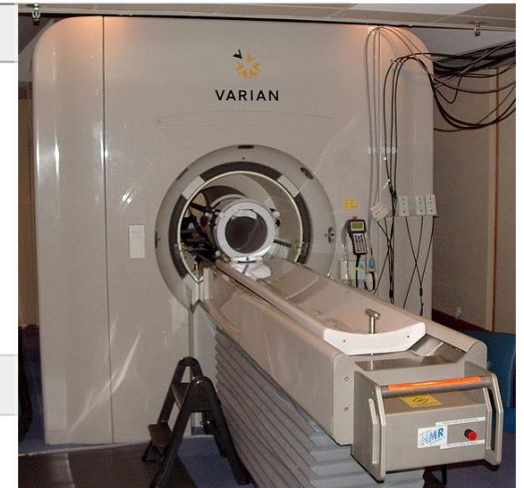


Published February 07, 2007 by:
[Andrea Okrentowich](#)
[View Profile](#) | [Follow](#) | [Add to Favorites](#)

More: [Brain Scans](#) | [Brain Scan](#) | [Disposition](#) | [Defendant](#) | [Criminal Behavior](#)

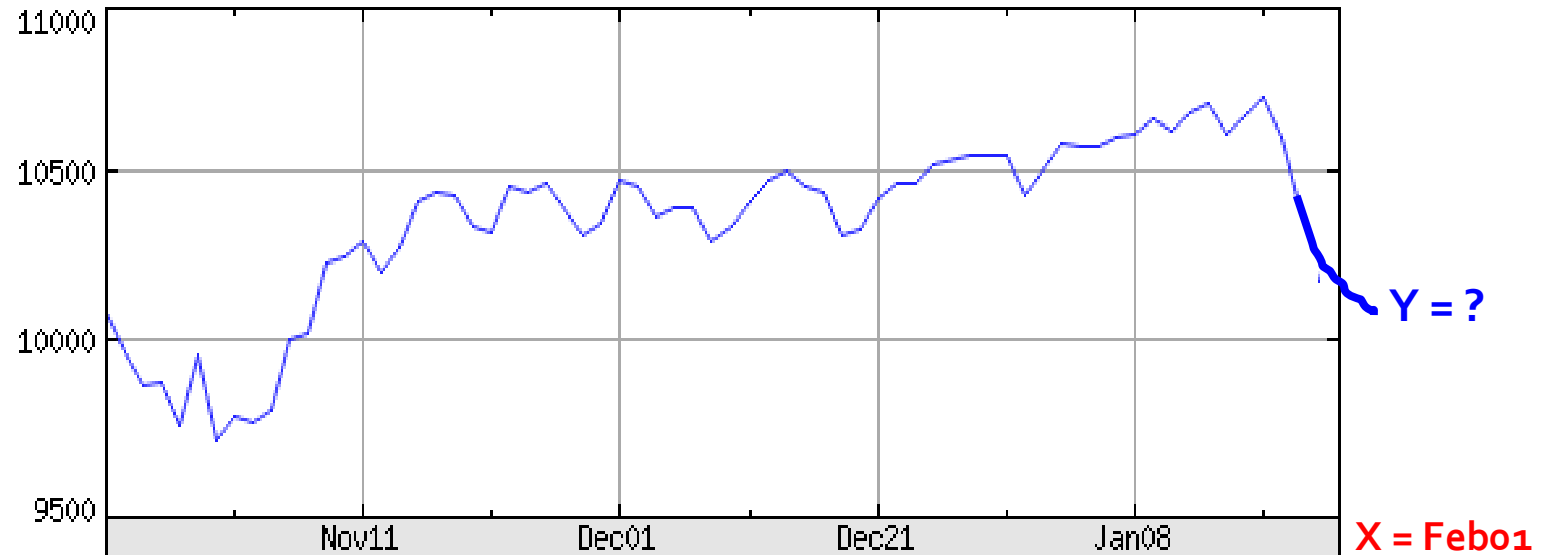
MRI Scans as Courtroom Evidence

The average Joe's MRI scan can show a brain abnormality, do we proceed to check him into the nearest mental institution or prison? That would make about as much sense as trying to prove a defendant innocent of a violent



Machine Learning in Action

DJ INDU AVERAGE (DOW JONES & CO
as of 22-Jan-2010



Copyright 2010 Yahoo! Inc.

<http://finance.yahoo.com/>

Machine Learning in Action

- Cars navigating on their own



Boss, the self-driving SUV 1st place in the DARPA Urban Challenge. Photo courtesy of Tartan Racing.



Machine Learning in Action

- The **best** helicopter pilot is now a computer!
 - it runs a program that learns how to fly and make acrobatic maneuvers by itself!
 - no taped instructions, joysticks, or things like that ...



[<http://heli.stanford.edu/>]


But ...

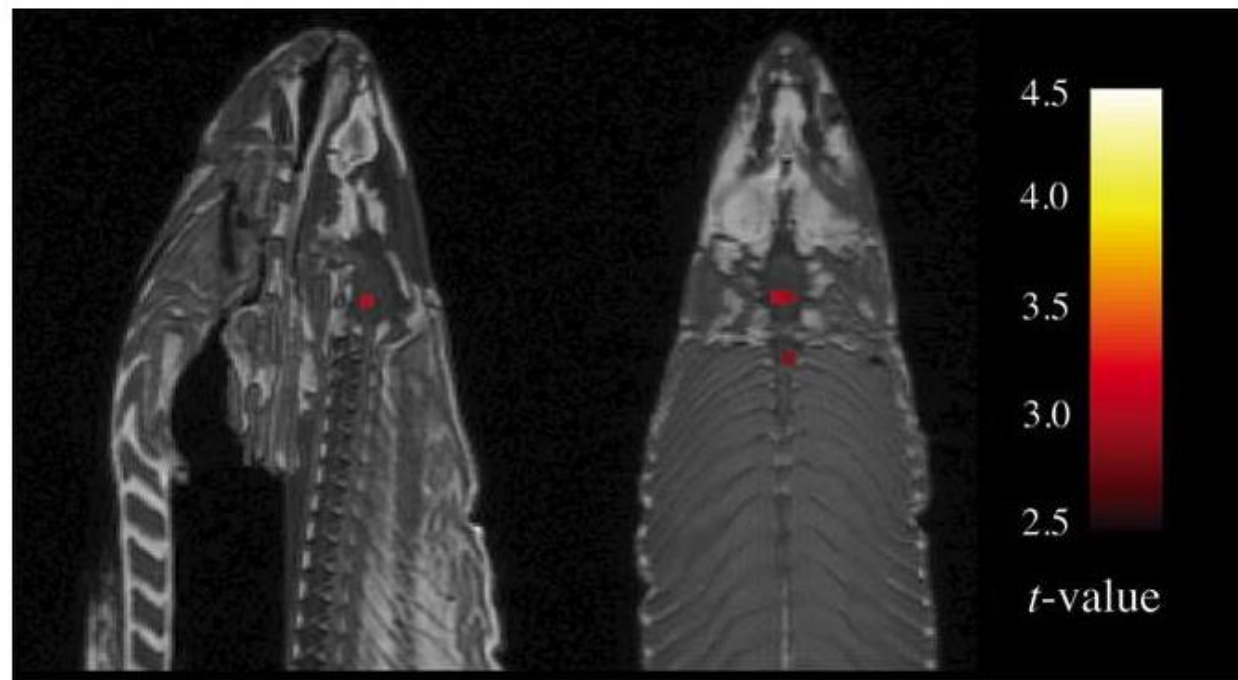
WIRED SCIENCE

NEWS FOR YOUR NEURONS



Scanning Dead Salmon in fMRI Machine Highlights Risk of Red Herrings

By Alexis Madrigal  September 18, 2009 | 5:37 pm | Categories: Brains and Behavior



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 institutions are accountable even when alleged discrimination is unintentional.

Misinformation on coronavirus is proving highly contagious

By DAVID KLEPPER July 29, 2020

The Switch

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



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.

ACLU

BECOME A MEMBER / RENEW / TAKE ACTION

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 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 Unbiased w/ Respect to Groups
 - Transparent
 - Interpretable
 - Robust to Adversarial Attack
 - Private
 - Environmentally Friendly
 - ⋮

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

“A Chinese woman [surname Yan] was offered two 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 ... It is unclear whether she purchased a third phone”

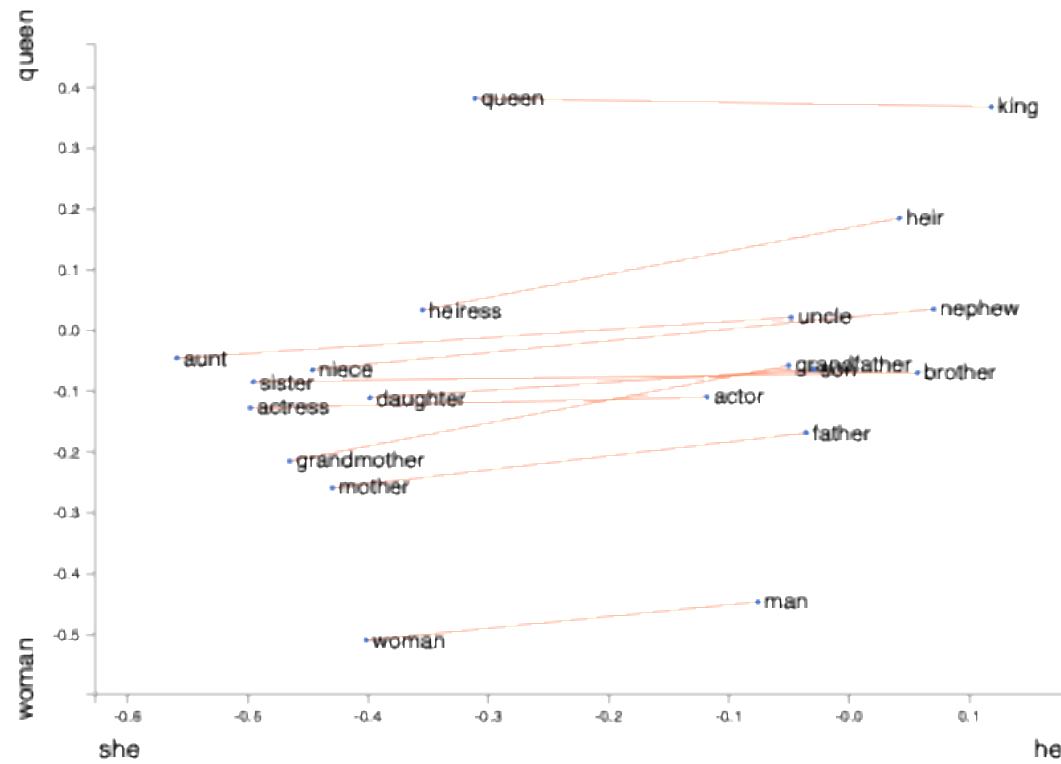
“As facial recognition systems become more common, Amazon has emerged as a frontrunner in the field, courting customers around the US, including police departments and Immigration and Customs Enforcement (ICE).”

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

Word embeddings and analogies



<https://lamiyowce.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?

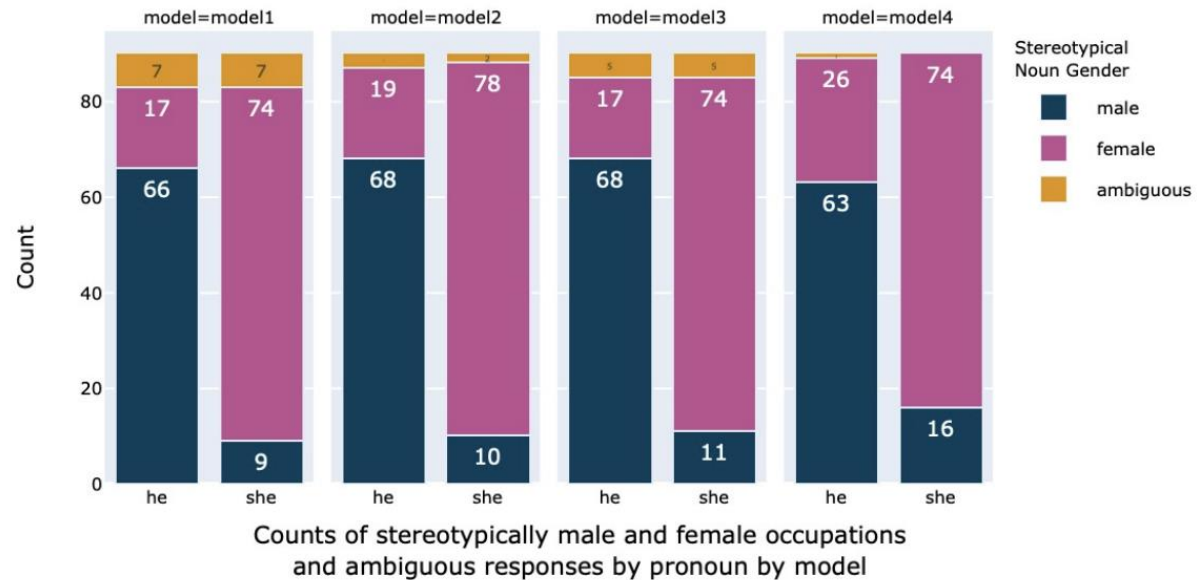


Figure 1: Occupation choices broken down by pronoun for the four models. Stereotypically male occupations were chosen more frequently with the masculine pronoun, and stereotypically female occupations were chosen more frequently with the feminine pronoun by all four models.

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."

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

May 23, 2016

Two Drug Possession Arrests



DYLAN FUGETT

LOW RISK

3



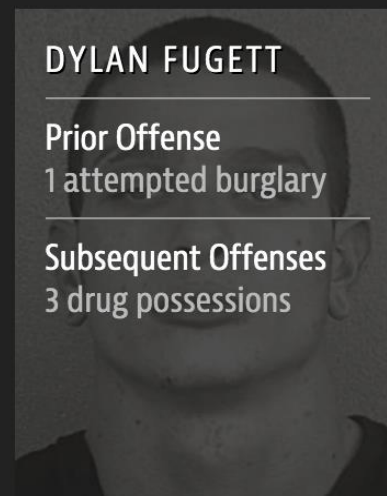
BERNARD PARKER

HIGH RISK

10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Two Drug Possession Arrests



DYLAN FUGETT

Prior Offense

1 attempted burglary

Subsequent Offenses

3 drug possessions

LOW RISK

3

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.



BERNARD PARKER

Prior Offense

1 resisting arrest
without violence

Subsequent Offenses

None

HIGH RISK

10

Different Types of Errors

	True label	Predicted label
True positive (TP)	+1	+1
False positive (FP)	-1	+1
True negative (TN)	-1	-1
False negative (FN)	+1	-1

Different Types of Performance Metrics

- Thus far, for binary classification tasks, we have largely only been concerned with accuracy i.e., minimizing the 0-1 loss
- Accuracy can be problematic in settings with...
 - Imbalanced labels e.g.,
 - Asymmetric costs for different types of errors e.g.,
- Two common alternatives are
 - Precision = # of true positives / # of predicted positives
= # of true positives / (# of true positives + # of false positives)
 - Recall = # of true positives / # of actual positives
= # of true positives / (# of true positives + # of false negatives)

F-score

- The F-score (or F_1 -score) of a classifier is the harmonic mean of its precision and recall:

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

How We Analyzed the COMPAS Recidivism Algorithm

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

May 23, 2016

	All Defendants		Black Defendants		White Defendants			
	Low	High	Low	High	Low	High		
Survived	2681	1282	Survived	990	805	Survived	1139	349
Recidivated	1216	2035	Recidivated	532	1369	Recidivated	461	505
FP rate: 32.35			FP rate: 44.85			FP rate: 23.45		
FN rate: 37.40			FN rate: 27.99			FN rate: 47.72		

False negative rate = $1 - \text{recall}$

This is one possible definition of unfairness.

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

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., $A \in \{a_1, \dots, a_c\}$
- $h(X, A) \in \{+1, -1\}$ 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
- 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.”

MythBusters #1

“If I remove the protected attribute A, I’m done.”



Myth: fairness through unawareness is enough.



Reality: other features X can still carry information about A through correlations or proxies.

A model that never sees A can still reproduce group disparities if X depends on A.



Key point: dropping A does not remove its footprint from the data.

Three Definitions of Fairness

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

Three Definitions of Fairness

- Independence: $h(X, A) \perp A$
- Separation:
- Sufficiency:

Independence

- Probability of being accepted 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,

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

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

Achieving Fairness

- Pre-processing data
- Additional constraints during training
- 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
	-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

- Probability of being accepted 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 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

MythBusters #2

“Equal acceptance rates means the system is fair.”



Myth: same selection rate across groups automatically means fairness.



Reality: that criterion is Independence. It can be satisfied by a lazy classifier that always predicts +1, or by random decisions for some groups.

Group A

60% admitted

model-based decisions



Group B

60% admitted

random decisions



Same rate



same quality

Three Definitions of Fairness

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

Three Definitions of Fairness

- **Independence:** $h(X, A) \perp A$
 - Probability of being accepted is the same for all genders
 - Permits laziness/is susceptible to adversarial decisions
- **Separation:** $h(X, A) \perp A \mid Y$
- **Sufficiency:**

Separation

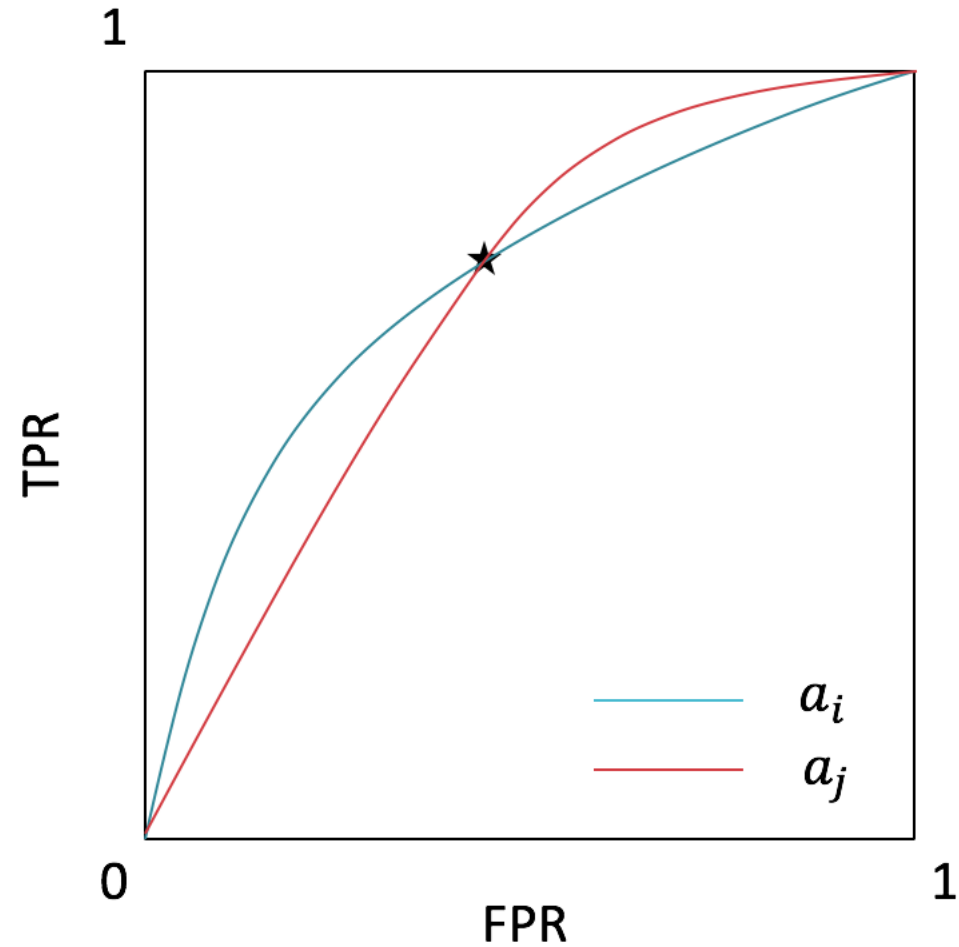
- Predictions and protected features can be correlated to the extent justified by the (latent) target variable

$$\begin{aligned} & P(h(X, A) = +1 | Y = +1, A = a_i) \\ &= P(h(X, A) = +1 | Y = +1, A = a_j) \text{ \& } \\ & P(h(X, A) = +1 | Y = -1, A = a_i) \\ &= P(h(X, A) = +1 | Y = -1, A = a_j) \forall a_i, a_j \end{aligned}$$

or equivalently, the model's true positive rate (TPR), $P(h(X, A) = +1 | Y = +1)$, 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

Achieving Separation



- ROC curve plots TPR against FPR at different prediction thresholds, τ :
 $h(X, A) = \mathbb{1}(\text{SCORE} \geq \tau)$
- 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

$$\begin{aligned} & P(h(X, A) = +1 | Y = +1, A = a_i) \\ &= P(h(X, A) = +1 | Y = +1, A = a_j) \text{ \& } \\ & P(h(X, A) = +1 | Y = -1, A = a_i) \\ &= P(h(X, A) = +1 | Y = -1, A = a_j) \forall a_i, a_j \end{aligned}$$

or equivalently, the model's true positive rate (TPR), $P(h(X, A) = +1 | Y = +1)$, 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:** $h(X, A) \perp A$
 - Probability of being accepted is the same for all genders
 - Permits laziness/is susceptible to adversarial decisions
- **Separation:** $h(X, A) \perp A \mid Y$
 - 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:** $h(X, A) \perp A$
 - Probability of being accepted is the same for all genders
 - Permits laziness/is susceptible to adversarial decisions
- **Separation:** $h(X, A) \perp A \mid Y$
 - All “good” applicants are accepted with the same probability, regardless of gender
 - Perpetuates existing biases in the training data
- **Sufficiency:** $Y \perp A \mid h(X, A)$

Sufficiency

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

$$\begin{aligned} & P(Y = +1 | h(X, A) = +1, A = a_i) \\ &= P(Y = +1 | h(X, A) = +1, A = a_j) \ \& \\ & P(Y = +1 | h(X, A) = -1, A = a_i) \\ &= P(Y = +1 | h(X, A) = -1, A = a_j) \ \forall a_i, a_j \end{aligned}$$

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

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

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:** $h(X, A) \perp A$
 - Probability of being accepted is the same for all genders
 - Permits laziness/is susceptible to adversarial decisions
- **Separation:** $h(X, A) \perp A \mid Y$
 - All “good”/”bad” applicants are accepted with the same probability, regardless of gender
 - Perpetuates existing biases in the training data
- **Sufficiency:** $Y \perp A \mid h(X, A)$
 - Equally calibrated across groups
 - 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
Clearly model	Sufficiency	Equivalent
Conditional use accuracy	Sufficiency	Equivalent
Predictive parity	Sufficiency	Relaxation
Calibration within groups	Sufficiency	Equivalent
Darlington criterion (1), (2)	Sufficiency	Relaxation

Quick Game

Quick game: name that fairness notion

Match each description to Independence, Separation, or Sufficiency.

A Same overall acceptance rate across groups

B Same TPR and FPR across groups

C The same score means the same outcome likelihood across groups

Quick Game

Quick game: name that fairness notion

Match each description to Independence, Separation, or Sufficiency.

A Same overall acceptance rate across groups

B Same TPR and FPR across groups

C The same score means the same outcome likelihood across groups

Reveal: A = Independence, B = Separation, C = Sufficiency.

Three Definitions of Fairness

- **Independence:** $h(X, A) \perp A$
 - Probability of being accepted is the same for all genders
 - Permits laziness/is susceptible to adversarial decisions
- **Separation:** $h(X, A) \perp A \mid Y$
 - All “good”/”bad” applicants are accepted with the same probability, regardless of gender
 - Perpetuates existing biases in the training data
- **Sufficiency:** $Y \perp A \mid h(X, A)$
 - Equally calibrated across groups
 - For the purposes of predicting Y , the information contained in $h(X, A)$ is “sufficient”, A becomes irrelevant

Three Incompatible Definitions of Fairness

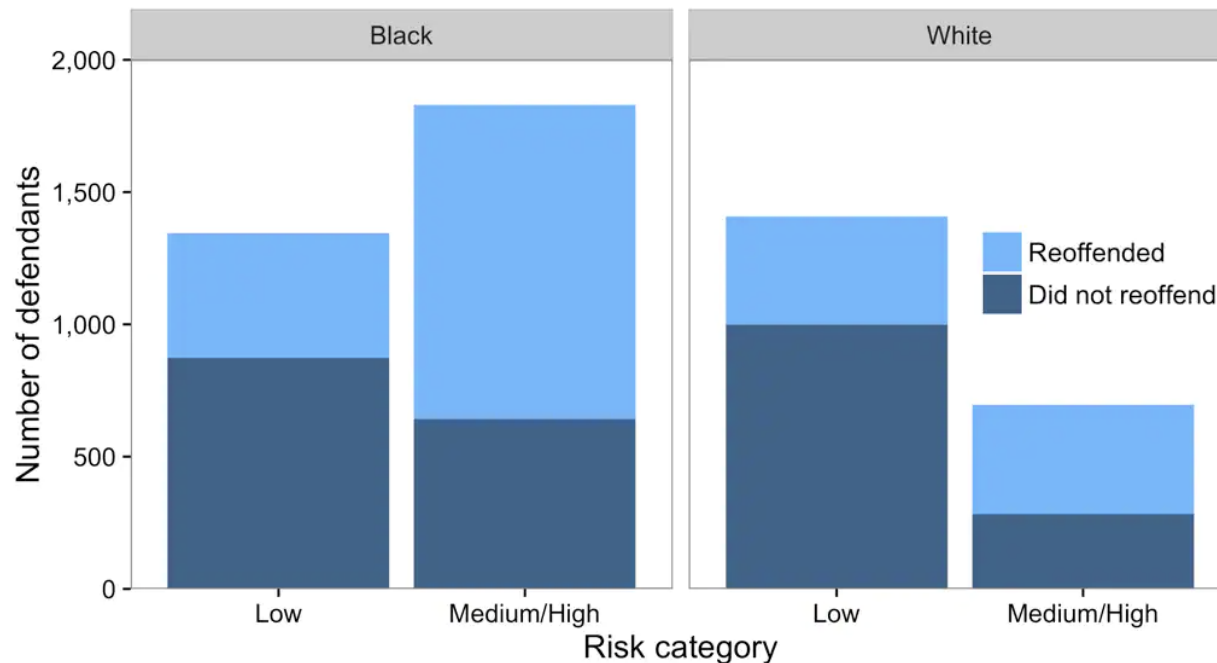
- **Independence:** $h(X, A) \perp A$
 - Probability of being accepted is the same for all genders
 - Permits laziness/is susceptible to adversarial decisions
- **Separation:** $h(X, A) \perp A \mid Y$
 - All “good”/”bad” applicants are accepted with the same probability, regardless of gender
 - Perpetuates existing biases in the training data
- **Sufficiency:** $Y \perp A \mid h(X, A)$
 - 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.

Quick Game

Two truths and a lie (fairness edition)

Vote for the lie: A, B, or C.

A. Separation is equivalent to equal TPR and FPR across groups.

B. Sufficiency is closely tied to calibration across groups.

C. With enough data, Independence, Separation, and Sufficiency can usually all be satisfied together.

Quick Game

Two truths and a lie (fairness edition)

Vote for the lie: A, B, or C.

A. Separation is equivalent to equal TPR and FPR across groups.

B. Sufficiency is closely tied to calibration across groups.

C. With enough data, Independence, Separation, and Sufficiency can usually all be satisfied together.

Reveal: C is the lie.

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
 - Independence: $h(X, A) \perp A$
 - Separation: $h(X, A) \perp A \mid Y$
 - Sufficiency: $Y \perp A \mid h(X, A)$
 - In all but the simplest of cases, any two of these three are mutually exclusive



Scan for poll:

Lecture 14

Final Poll

Final poll (pick one)

Which statement best matches the lecture's main takeaway about Independence, Separation, and Sufficiency?

- A. Removing A from the input is enough to guarantee fairness.
- B. Independence, Separation, and Sufficiency are basically the same notion.
- C. They capture different goals, and in all but the simplest cases any two of the three are mutually exclusive.
- D. Separation means equal overall acceptance rates across groups.
- E. Sufficiency says TPR and FPR must be equal across groups.