DATA SCIENCE
INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

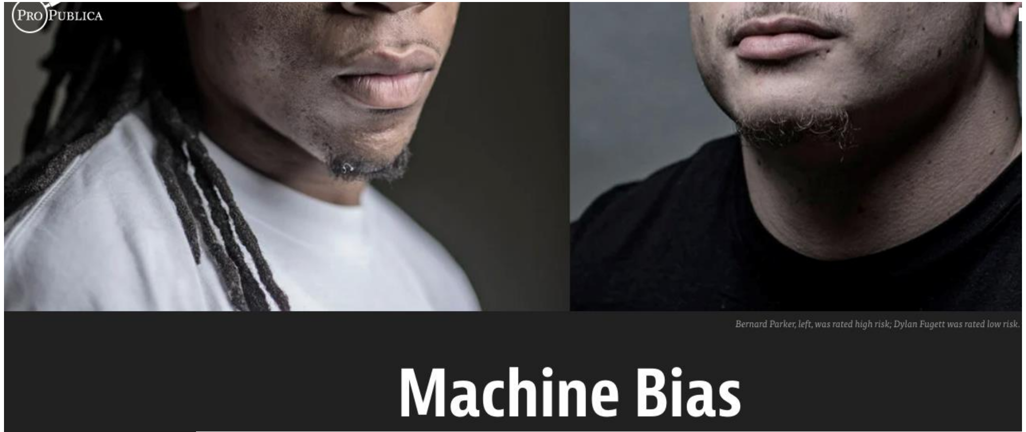
Bias in AI: toward building fair and equitable healthcare applications

Monica J. Wood, MD
Department of Radiology, Massachusetts General Hospital

I have no relevant conflict of interest to disclose.

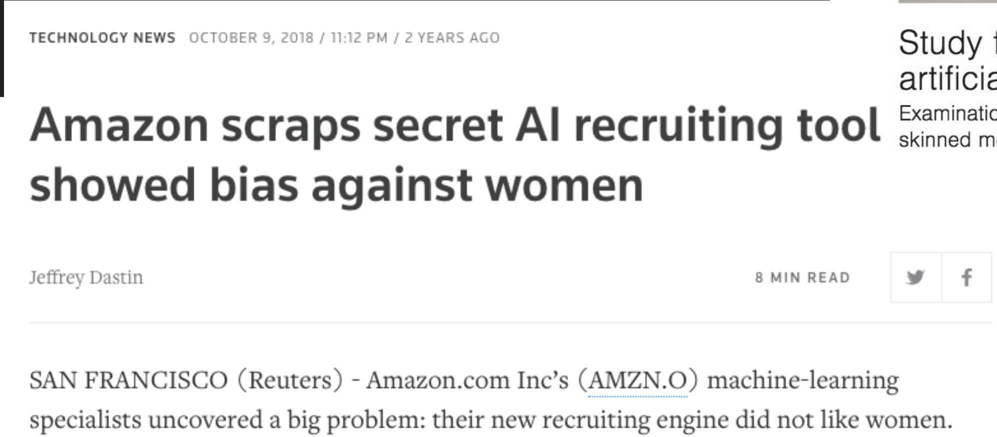
Humans are biased

But can machines be biased too?



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk.

Machine Bias



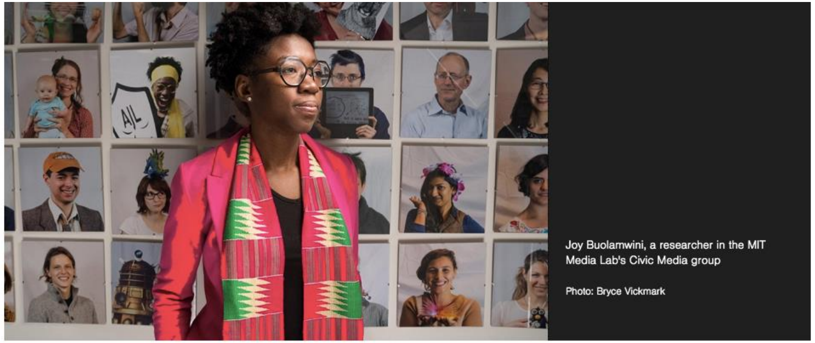
TECHNOLOGY NEWS OCTOBER 9, 2018 / 11:12 PM / 2 YEARS AGO

Amazon scraps secret AI recruiting tool showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.



Joy Buolamwini, a researcher in the MIT Media Lab's Civic Media group

Photo: Bryce Vickmark

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

@monicajwood

1) How do biases make their way into ML algorithms?

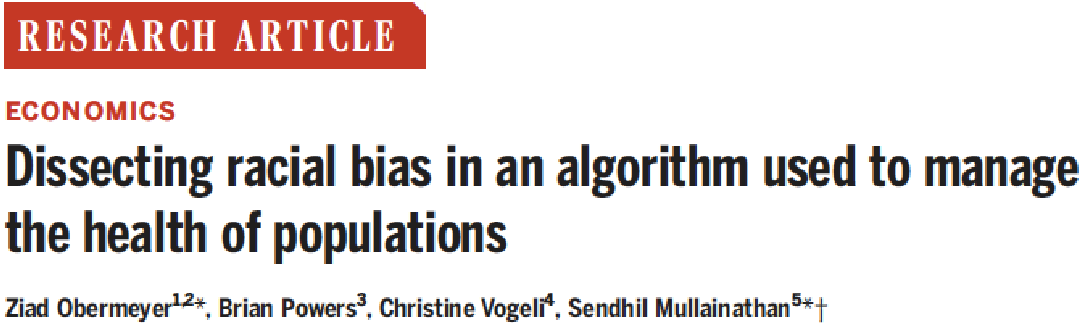
2) How do we minimize bias and strive for fairness in AI applications?

3) How can the ML/AI community build fair and equitable healthcare applications?

How do biases make their way into ML algorithms?

Sources of bias: training data

- Training data may include the result of biased human decisions or the effects of historical or systemic inequities



RESEARCH ARTICLE

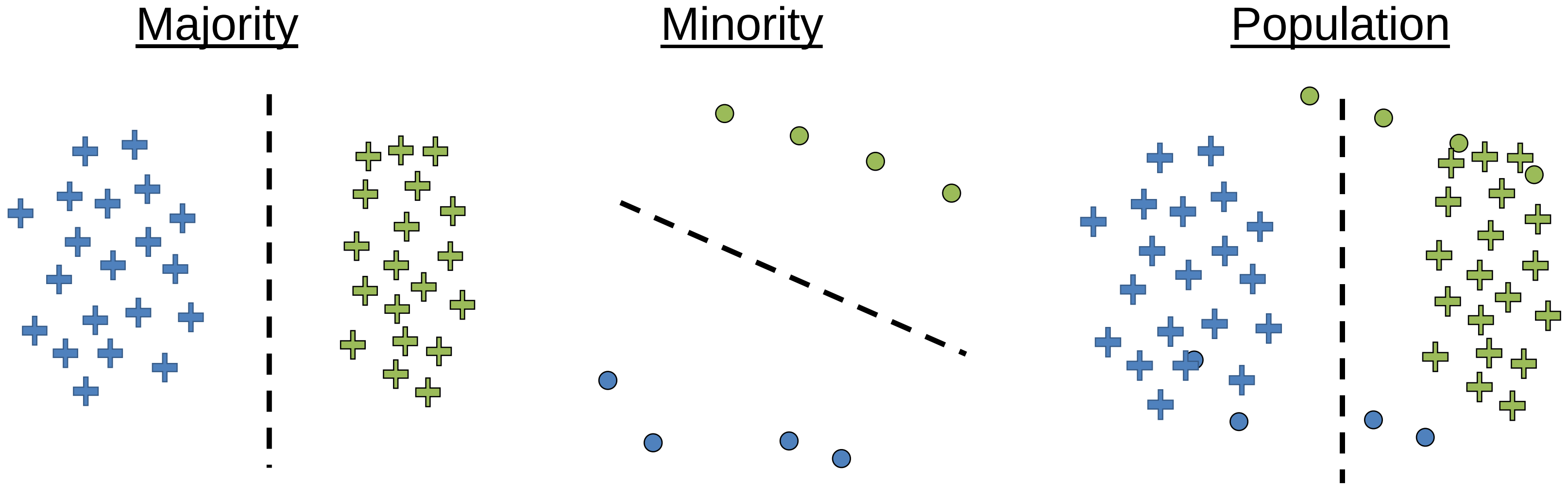
ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Sources of bias: training data

- Under-representation of a sub-population in the dataset may result in decreased performance of the trained model



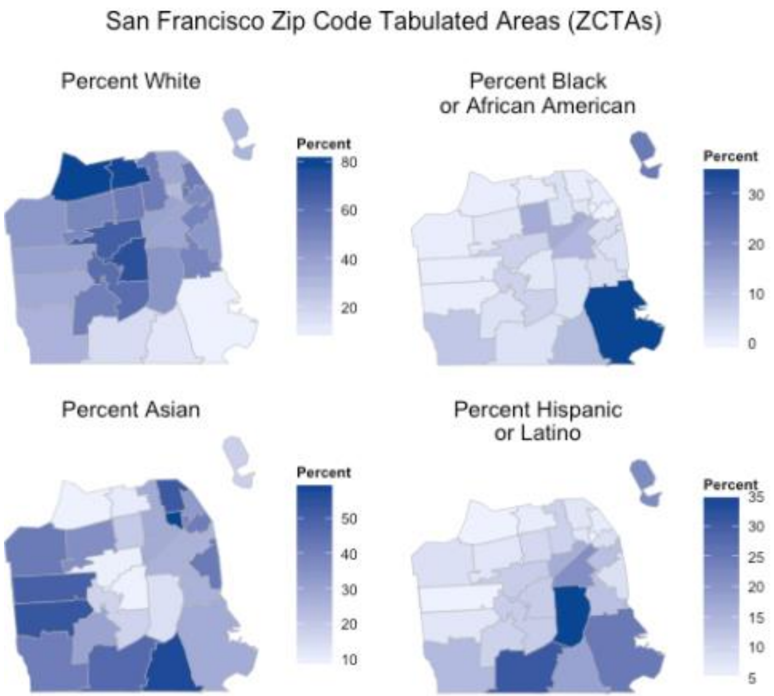
Majority

Minority

Population

Sources of bias: training data

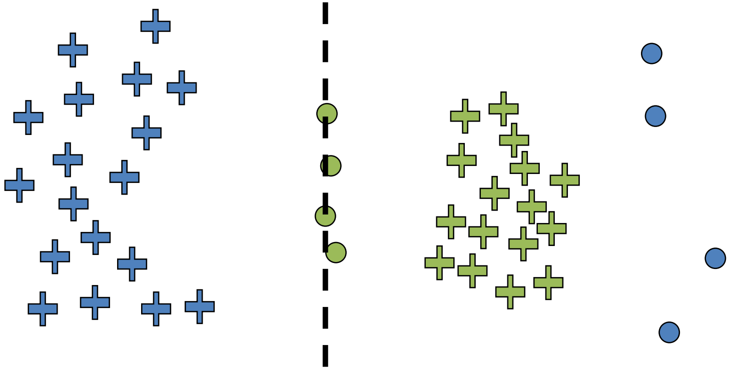
- Masked variables may remain present in the dataset through correlates (e.g., race and zip code)


San Francisco Zip Code Tabulated Areas (ZCTAs)
Percent White
Percent
80
60
40
20
Percent Black or African American
Percent
30
20
10
0
Percent Asian
Percent
50
40
30
20
10
Percent Hispanic or Latino
Percent
35
30
25
20
15
10
5

Source: https://blog.revolutionanalytics.com/2015/04/exploring-san-francisco-with-choropleth.html

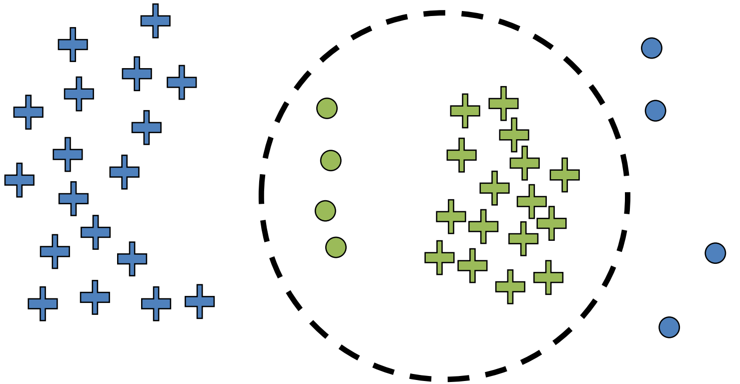
Sources of bias: algorithm design

- The type of ML architecture or variables chosen can favor the majority sub-population at the detriment of a minority sub-population



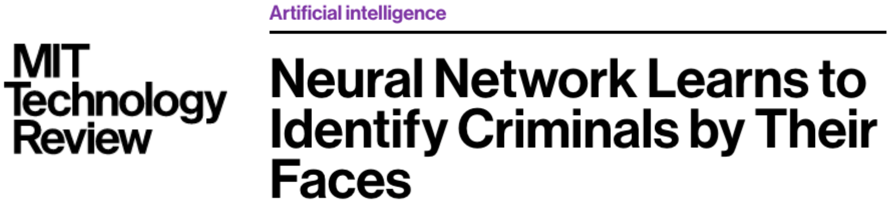
Sources of bias: algorithm design

- The type of ML architecture or variables chosen can favor the majority sub-population at the detriment of a minority sub-population



Sources of bias: model output and application

- Human actions based upon biased model output may perpetuate existing bias
- Positive feedback loops may amplify existing biases
- Applications may be used for discriminatory purposes



MIT Technology Review

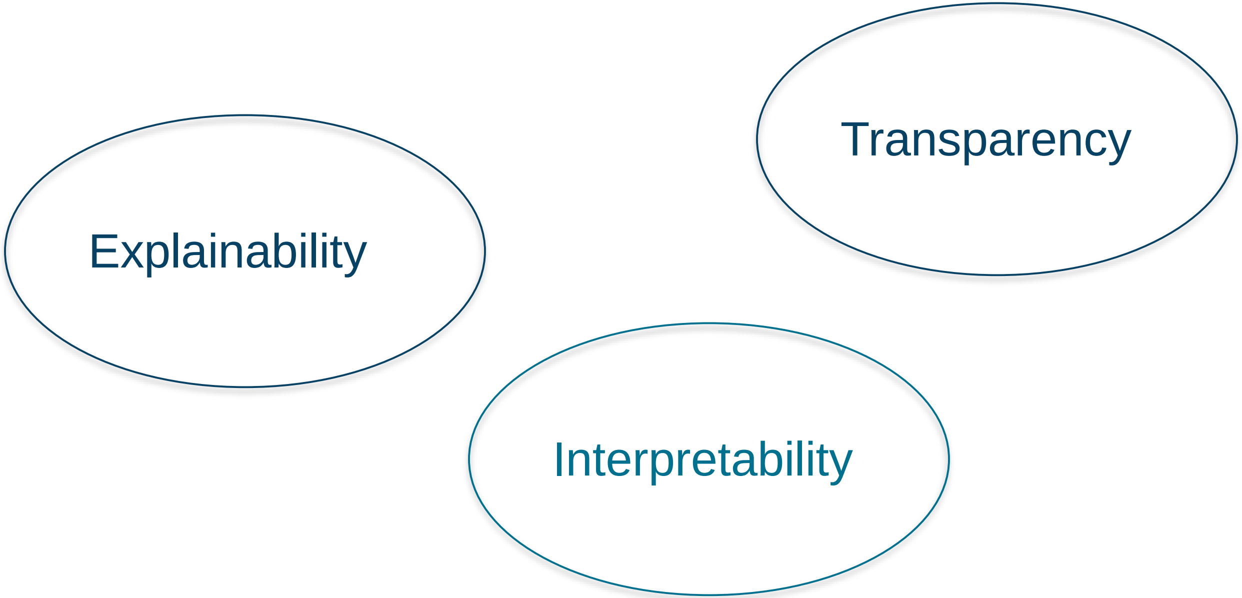
Artificial intelligence

Neural Network Learns to Identify Criminals by Their Faces

How do we minimize bias

and strive for fairness in AI applications?

Understanding ML algorithms



Transparency

Explainability

Interpretability

Defining and measuring fairness

- ✓ Defining fairness and establishing metrics to assess fairness are very challenging tasks

- ✓ Trade-offs: a given algorithm cannot necessarily satisfy multiple fairness metrics to achieve individual and group fairness along multiple axes

- ✓ Deciding on what is fair will require multidisciplinary expertise and collaboration

Addressing bias and fairness at every step

✓ Process the data to address biases before using for training

✓ Incorporate fairness definitions into the training process

✓ Scrutinize and even modify the outputs before operationalizing

Incorporating bias evaluation in QI/QA processes

- ✓ Check overall accuracy and by subgroup

- ✓ Consider ‘counterfactual fairness’
  - What would have happened if the patient had been of a different *gender/race/ethnicity*?

- ✓ Use domain knowledge to uncover when the majority solution may harm a minority sub-population

How can the ML/AI community move forward

building fair and equitable healthcare applications?

A call to action

- Commit to diversifying AI talent in healthcare: who creates, validates, and monitors models?

- Stay informed: Fairness, Accountability, and Transparency has emerged as a constantly evolving research field (**fatml.org**)

- Have the hard conversations: be explicit about an algorithm’s objectives and trade-offs

Summary

- Unwanted bias may be reflected in AI algorithms via the training data used, the model design selected, and the applications of the algorithm output

- Steps to mitigate bias include achieving a deeper understanding of how algorithms are constructed, agreeing on measurable and relevant definitions of fairness, and proactively evaluating for potential bias

- A diverse AI workforce engaged in promoting fairness, accountability, and transparency can pave the way toward building fair and equitable AI healthcare applications

Thank you

Monica J. Wood, MD
Department of Radiology
Massachusetts General Hospital

mwood9@mgh.harvard.edu

DATA SCIENCE INSTITUTE®
AMERICAN COLLEGE OF RADIOLOGY

@monicajwood