Responsible Machine Learning

Lecture 4: Bias in the Wild

CS 4973-05

Fall 2023

Instructor: Avijit Ghosh
ghosh.a@northeastern.edu
Northeastern University, Boston, MA

Northeastern University

Today's lecture credits: Data Crunch Lab, Yee et Al. @ Twitter, Vinodkumar Prabhakaran

@DataCrunch_Lab
@RTPAnalysts

Sources of Bias

(CC BY-SA 2.0) Photo by Chris Bloom

- ❑Skewed sample
- ❑Tainted examples
- ❑Sample size disparity
- ❑Limited features
- ❑Proxies

DATACRUNCH Lab

@DataCrunch_Lab
@RTPAnalysts

Skewed sample

Detect potholes to allocate repair crews

"The system reported a disproportionate number of potholes in wealthier neighbourhoods. It turned out it was oversampling the younger, more affluent citizens who were digitally clued up enough to download and use the app in the first place."

- David Wallace, "Big data has unconscious bias too", Sept. 21, 2016

DATACRUNCH Lab

@DataCrunch_Lab
@RTPAnalysts

Tainted examples

Identify job candidates

"That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry."

- Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women", Oct. 9, 2018

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

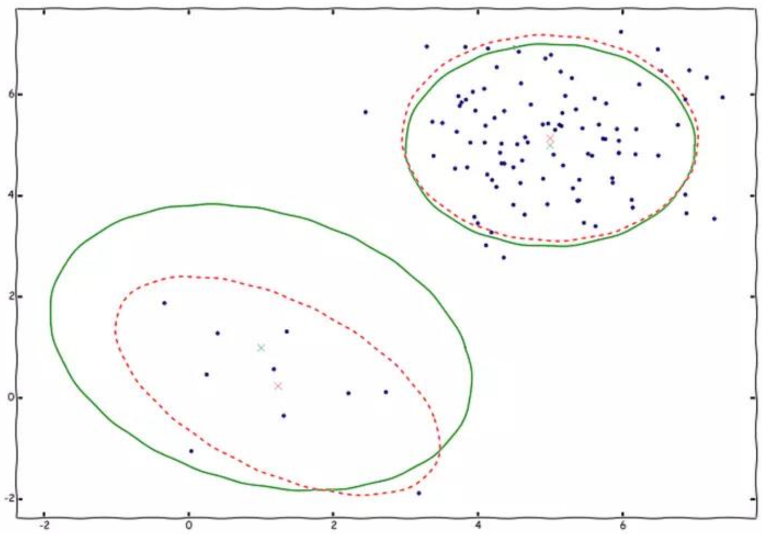
Jeffrey Dastin 8 MIN READ

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

DataCrunch Lab

@DataCrunch_Lab
@RTPAnalysts

Sample size disparity



"Assuming a fixed feature space, a classifier generally improves with the number of data points used to train it… The contrapositive is that less data leads to worse predictions. "

\- Moritz Hardt, "How big data is unfair", Sept. 26, 2014

DATACRUNCH Lab

@DataCrunch_Lab
@RTPAnalysts

Limited features

Diagnose heart disease

“recent research has shown that women who have heart disease experience different symptoms, causes and outcomes than men.”

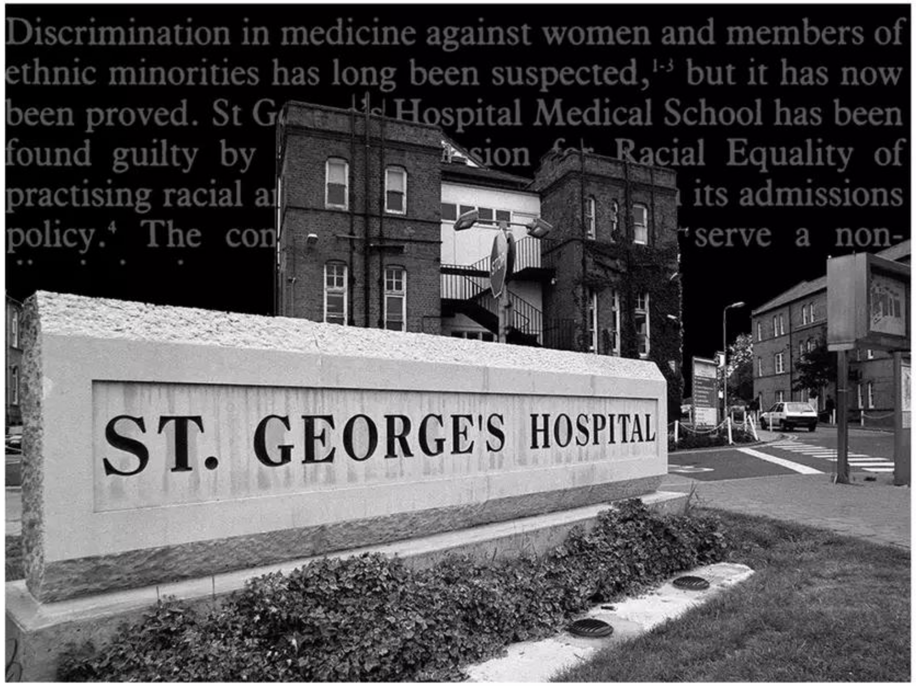
“Women are often told their stress tests are normal or that they have "false positives." Bairey Merz says doctors should pay attention to symptoms such as chest pain and shortness of breath rather than relying on a stress test score.”

DataCrunch Lab

@DataCrunch_Lab
@RTPAnalysts

Proxies

Make college admissions decisions



*"...certain rules in the system that weighed applicants on the basis of seemingly non-relevant factors, like **place of birth** and **name**.*

[...]simply having a non-European name could automatically take 15 points off an applicant's score. The commission also found that female applicants were docked three points, on average."

- Oscar Schwartz, "Untold History of AI: Algorithmic Bias was Born in the 1980s", April 15, 2019

@DataCrunch_Lab
@RTPAnalysts

PRO PUBLICA

Donate

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

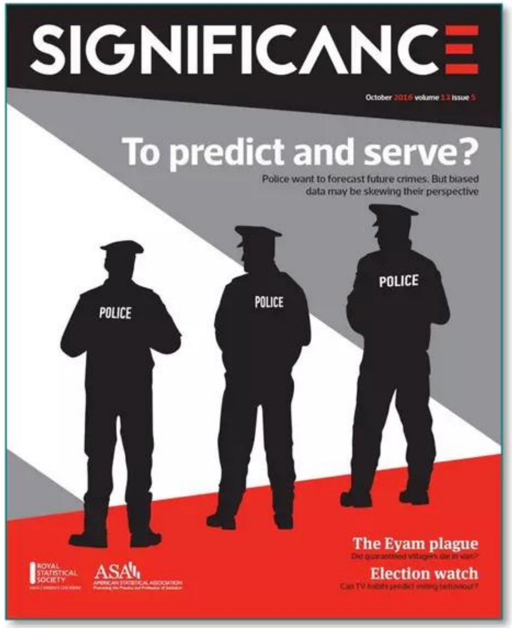
May 23, 2016

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

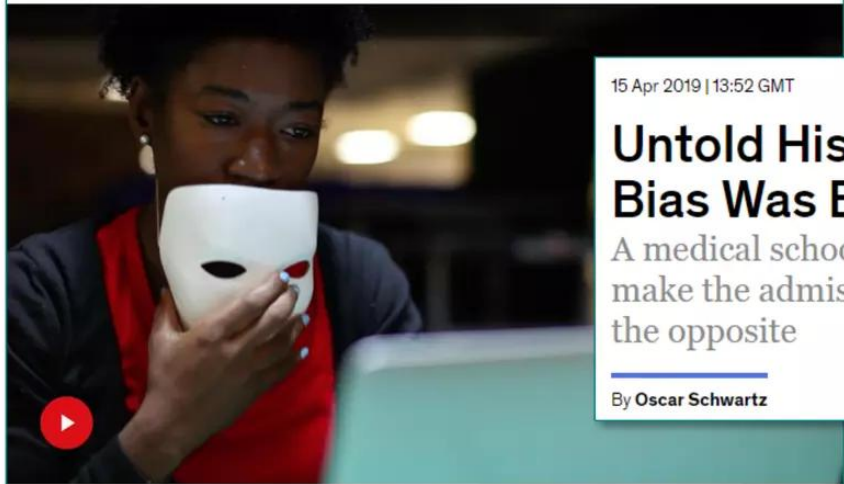
8 MIN READ



TIME

IDEAS • THE ART OF OPTIMISM

Artificial Intelligence Has a Problem With Gender and Racial Bias. Here's How to Solve It



BY JOY BUOLAMWINI FEBRUARY 7, 2019

IDEAS *Buolamwini is a computer scientist, founder of the Algorithmic Justice League and a poet of code.*

15 Apr 2019 | 13:52 GMT

Untold History of AI: Algorithmic Bias Was Born in the 1980s

A medical school thought a computer program would make the admissions process fairer—but it did just the opposite

By **Oscar Schwartz**

DATACRUNCH *Lab*

@DataCrunch_Lab
@RTPAnalysts

"Scores like this — known as risk assessments — are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts — as is the case in Fort Lauderdale — to even more fundamental decisions about defendants' freedom. In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing."

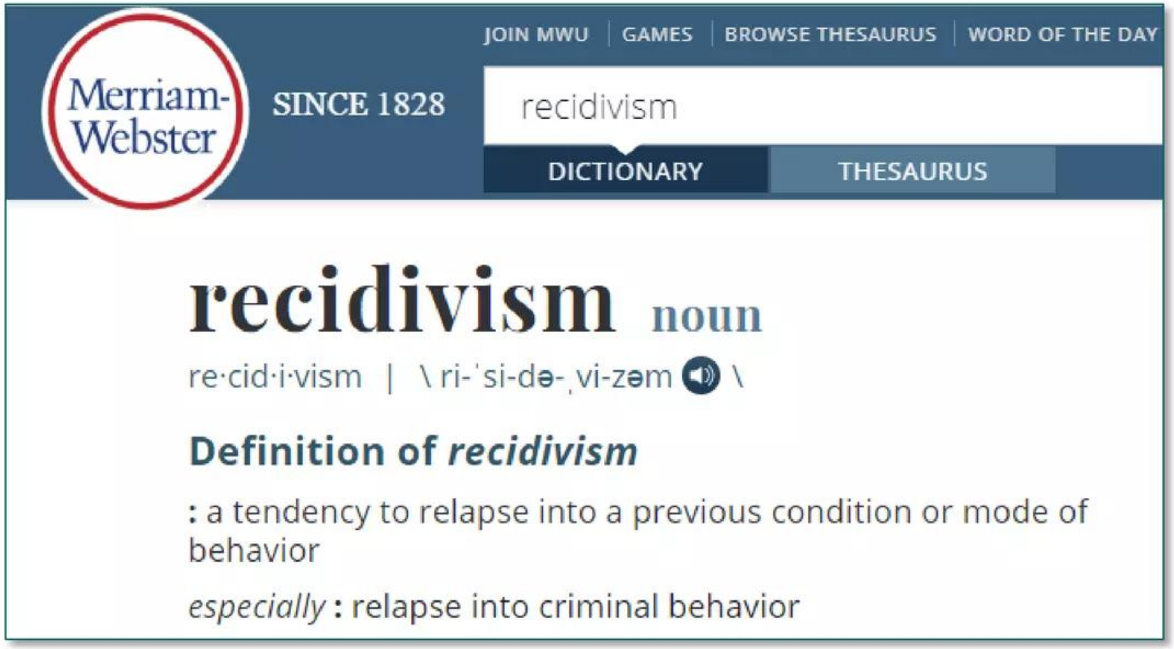
– Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, ProPublica, May 23, 2016

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

DataCrunch Lab

@DataCrunch_Lab
@RTPAnalysts

Recidivism



Northpointe

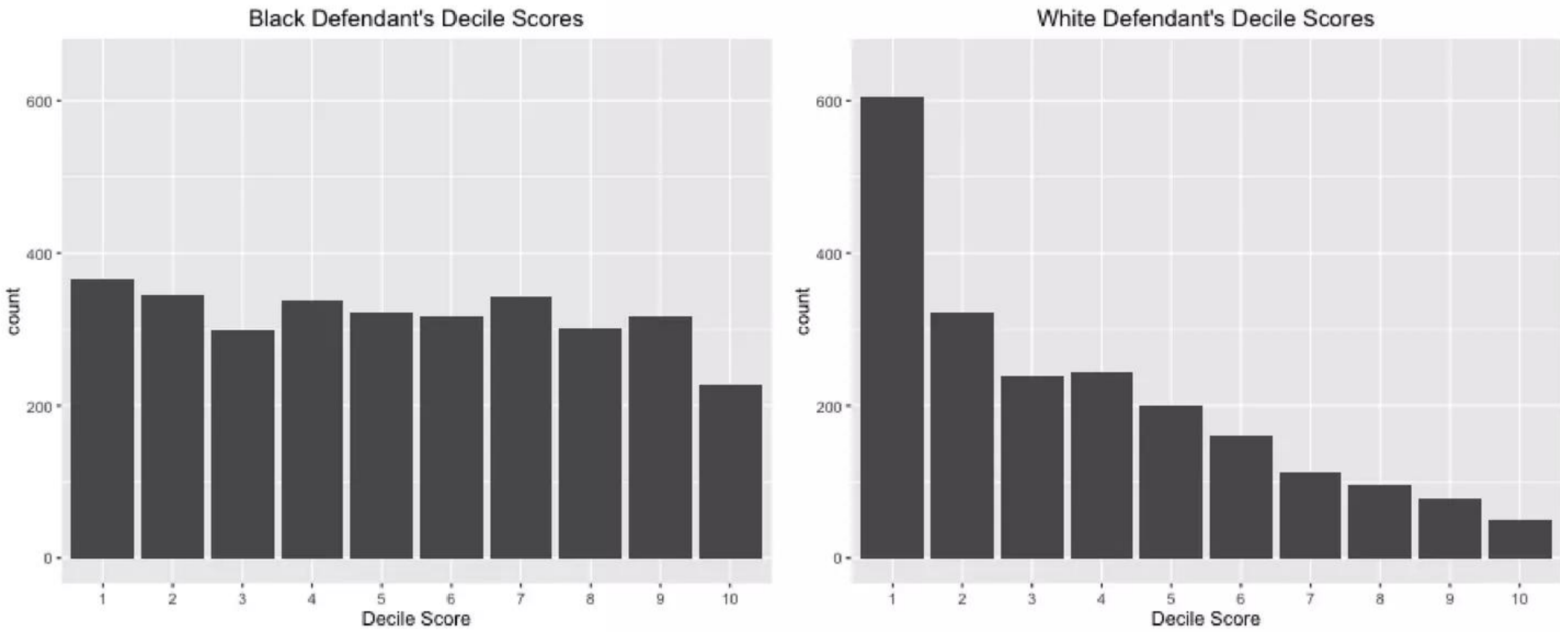
“a finger-printable ***arrest*** *involving a charge and a filing for any uniform crime reporting (UCR) code.”*

ProPublica

“criminal offense that resulted in a jail booking”

DataCrunch Lab

Risk of Recidivism Score



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

@DataCrunch_Lab
@RTPAnalysts

Analysis of COMPAS Recidivism Score

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

DATACRUNCH Lab

@DataCrunch_Lab
@RTPAnalysts

Accurate for all groups

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**

Accuracy

“ The company said it had devised the algorithm to achieve this goal. A test that is correct in equal proportions for all groups cannot be biased, the company said.”

All defendants	Black defendants	White defendants
65.4%	**63.8%**	**67.0%**

DATACRUNCH Lab

@DataCrunch_Lab
@RTPAnalysts

Black defendants wrongly classified high risk more often

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

Error rate

"Black defendants who do not recidivate were nearly twice as likely to be classified by COMPAS as higher risk compared to their white counterparts."

32.3%	**44.8%**	**23.5%**

DATACRUNCH Lab

@DataCrunch_Lab
@RTPAnalysts

White defendants wrongly classified low risk more often

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

Error rate

"The test tended to make the opposite mistake with whites, meaning that it was more likely to wrongly predict that white people would not commit additional crimes if released compared to black defendants."

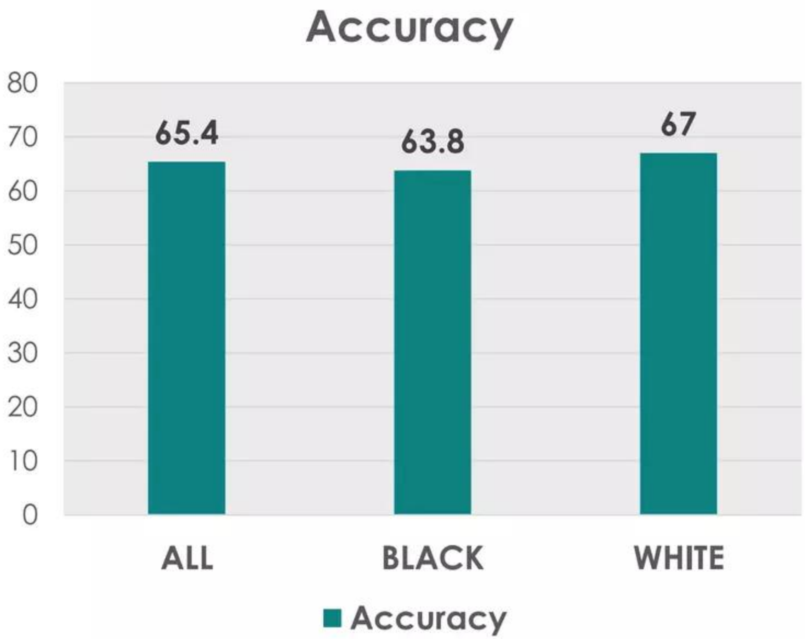
37.4% **28.0%** **47.7%**

DATACRUNCH Lab

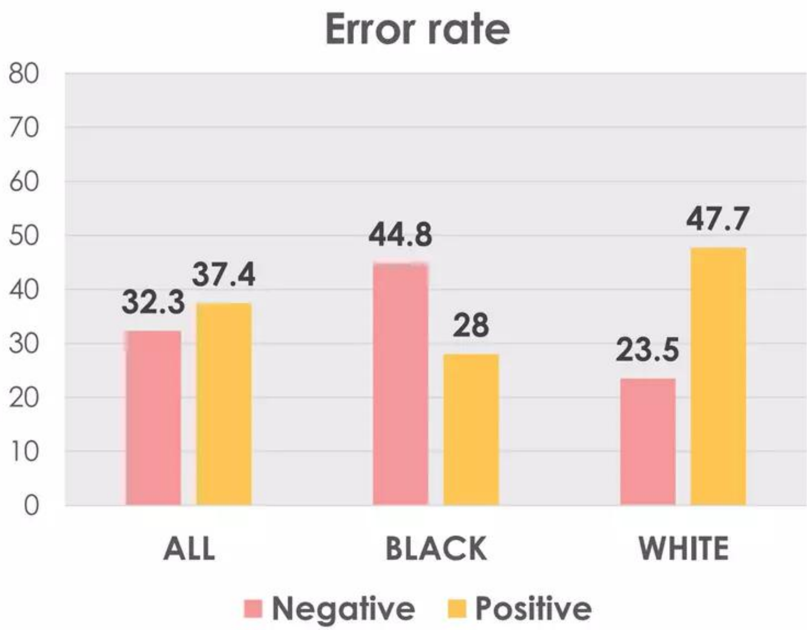
@DataCrunch_Lab
@RTPAnalysts

Biased or Unbiased?

Predictive Parity



Disparate Impact



DataCrunch Lab

HW Q1: The graphics in the article illustrate a tension between equalizing error rates across groups and choosing a single threshold for all people. Why was it impossible to achieve both of these at the same time?

@DataCrunch_Lab
@RTPAnalysts

"Since blacks are re-arrested more often than whites, is it possible to create a formula that is equally predictive for all races without disparities in who suffers the harm of incorrect predictions?"

NO

DataCrunch Lab

@DataCrunch_Lab
@RTPAnalysts

Key finding

(CC BY 2.0) Photo by Ivan Radic

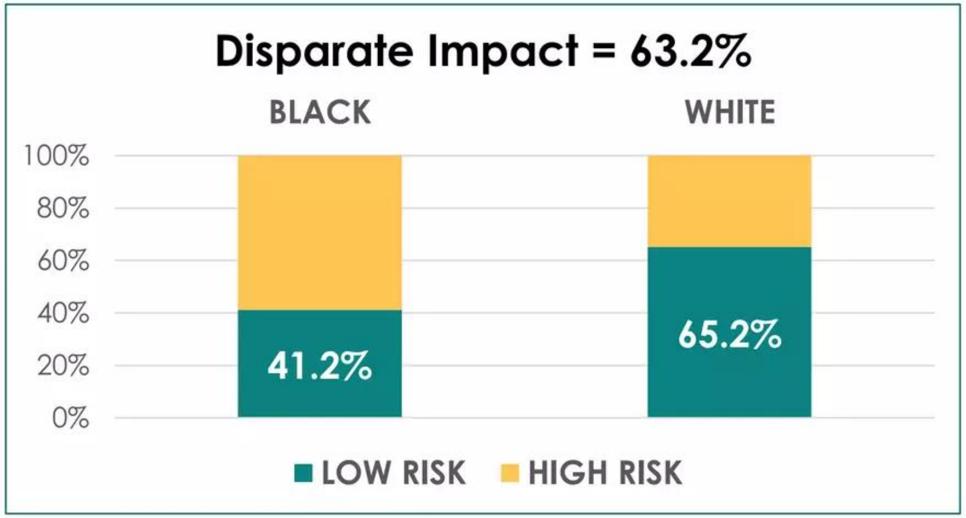
"If you have two populations that have unequal base rates then you can't satisfy both definitions of fairness at the same time."

@DataCrunch_Lab
@RTPAnalysts

Fairness Metrics

Disparate Impact

Ratio of the rate of favorable outcomes between the unprivileged and privileged groups



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

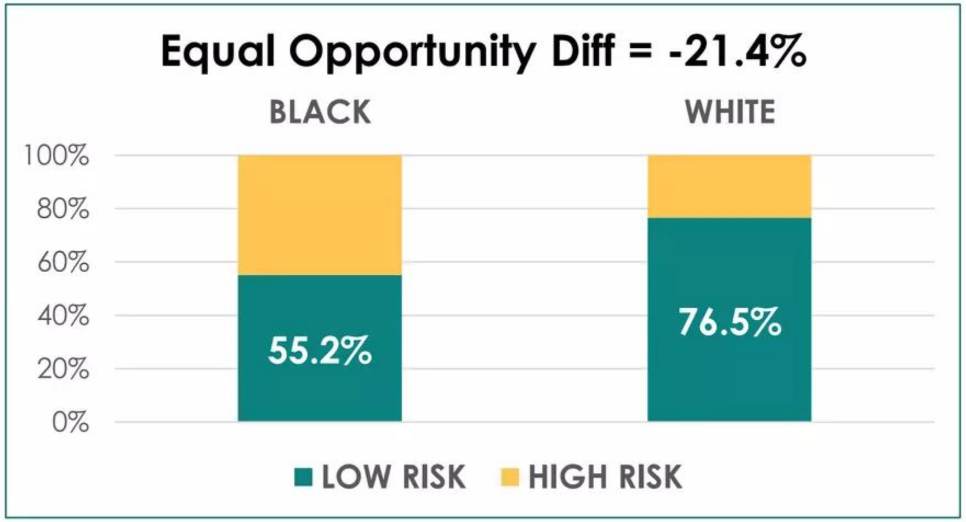
DATACRUNCH Lab

@DataCrunch_Lab
@RTPAnalysts

Fairness Metrics

Equal Opportunity Difference

Difference in true positive rates between the unprivileged and privileged groups



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

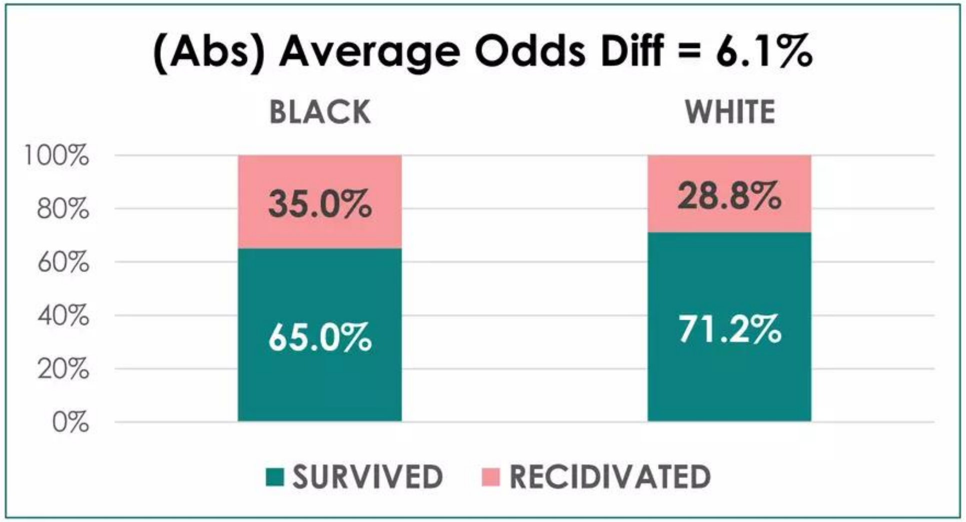
DATACRUNCH Lab

@DataCrunch_Lab
@RTPAnalysts

Fairness Metrics

Average Odds Difference

Average of difference in false positive rate and true positive rates of favorable outcomes between the unprivileged and privileged groups



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

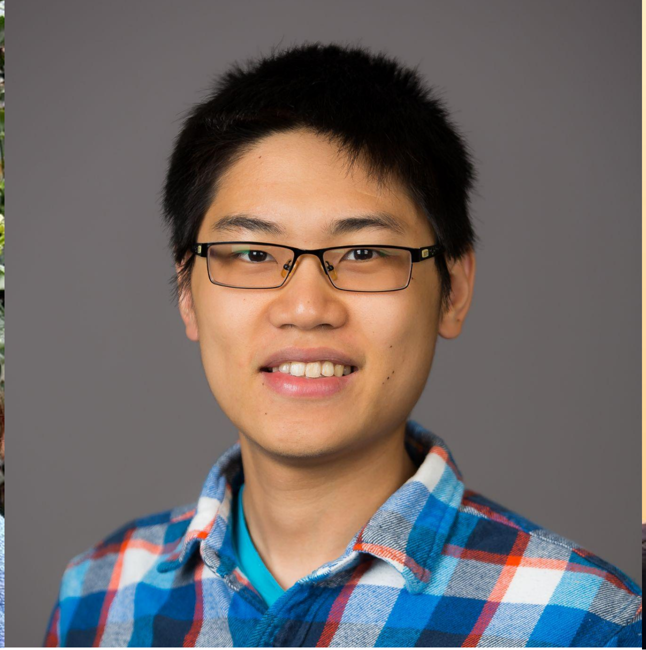
DATACRUNCH Lab

Image Cropping on Twitter:

Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency

Kyra Yee

ML Researcher, META team



Uthaipon (Tao) Tantipongpipat

ML Researcher, META team

Shubhanshu Mishra

ML Researcher, CAR (CUR)

Image cropping outline:
What is representational harm
What's cropping algorithm?
Problems and solutions

- Problems: demographic parity and male gaze
- Quantitative results
  - Argmax
- Qualitative analysis
- What we learned

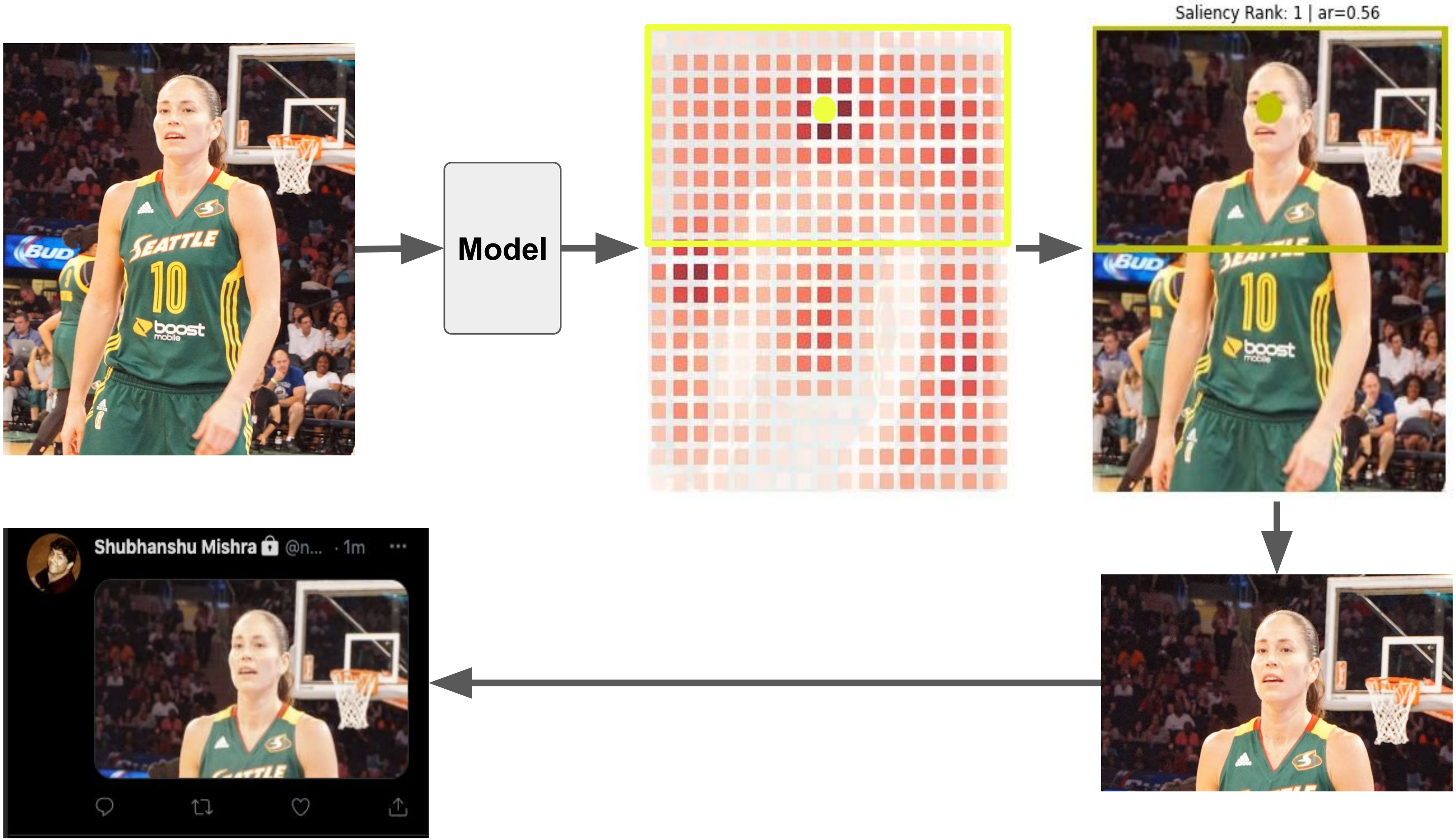
Representational Harm in Technology

- Allocative vs representational harms
- Representational harms lead to allocative harms
- People of color are simultaneously under and over exposed by technology - ex. Facial recognition

Image Cropping Algorithm

Task: original image + crop dimension ⇒ “best” (e.g. most important region) crop

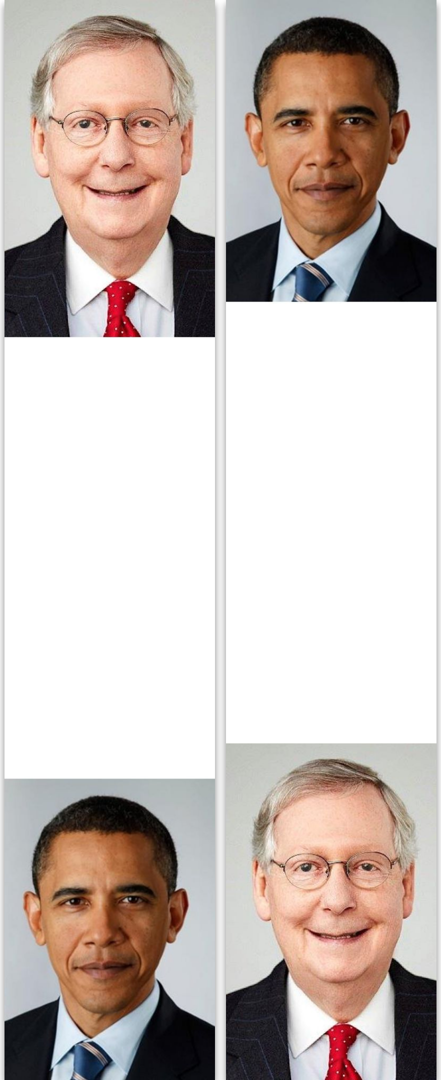
Example use: Image preview for difference devices (phone, laptop browser, etc.)


Model
Saliency Rank: 1 | ar=0.56
Shubhanshu Mishra @n... · 1m
SEATTLE
10
boost
mobile

HW Q2: What fairness metric did the Twitter team use to measure disparate impact? What is a non-technical interpretation of this metric?

Demographic Parity

- Images of two individuals are attached
- See which one the Twitter model crops in the preview
- Can appear as racist cropping



Male Gaze

- Images of women are cropped at the middle or bottom part of the body

Quantitative Analysis

Demographic Parity: How We Tested

- Collect public figure image, gender, and ethnicity
- Two tests: gender and ethnicity
- Split images into 4 subgroups:
    - Black-Female (group size = 621),
    - Black-Male (1,348),
    - White-Female (213), and
    - White-Male (606)

Maya Moore Red Team.jpg
1,448 × 1,862; 89 KB

sex or gender	female 1 reference
ethnic group	African Americans 0 references

Demographic Parity: How We Tested

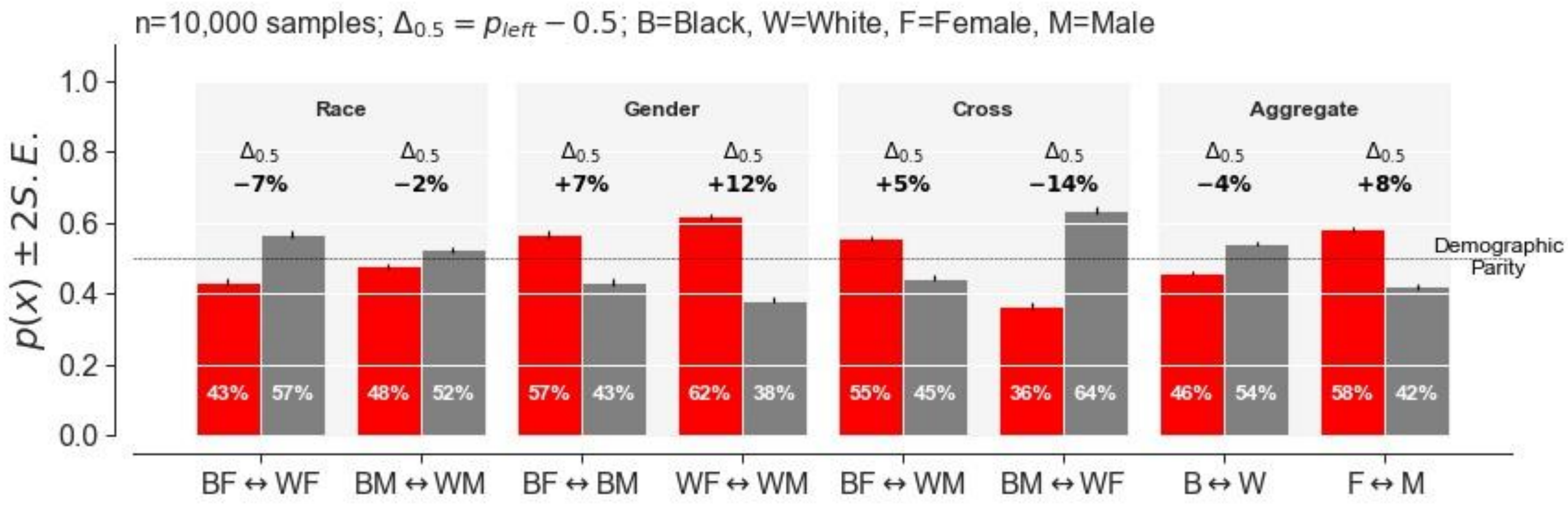
- Each pair of the subgroups → Sample one from each → Attach them
- Record which subgroup’s image (wherever it is) has the highest saliency
- Repeat sampling many times. 50-50 would be most equal.

$$\frac{P(R=1|A=a)}{P(R=1|A=b)} \leq 1-\epsilon$$

Black-Male > White-Male

“chosen”

HW Q3: Explain Figure 2. What is plotted? What do we observe?



95% confidence interval (after rounding) is ±1.0%

- Summary: Gender bias female > male is clear; Race bias is weaker.
  - Limitations exist from label; race and gender are more nuanced.

Limitations of Demographic Data

- Gender and race are not binary
- Given ethnic labels from wikipedia, we used US census race categories to standardize and simplify – Western centric analysis
- Race might not be the most suitable attribute to relate to images
- Risk of reifying racial and gender categories as natural rather than socially constructed - however, the goals is to study the impact on historically marginalized populations

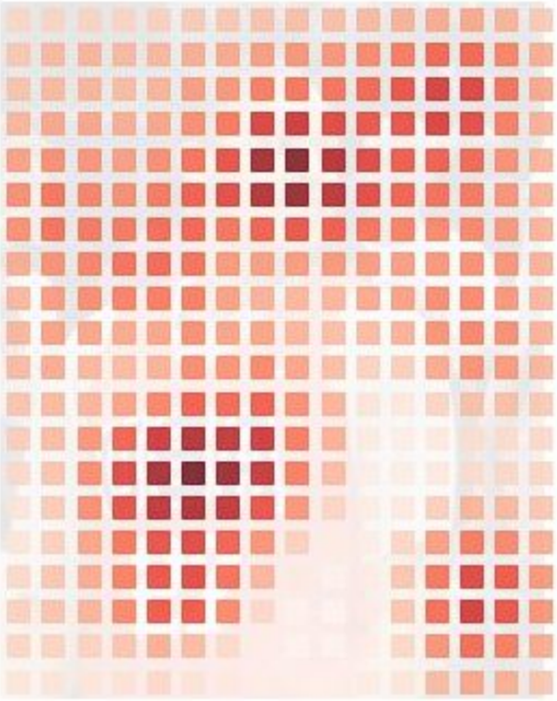
HW Q4: Explain Figure 5. What is plotted? What do we observe?

Male Gaze: What We Found

Spot checked 100 male and 100 female images with **>1** salient region.

Only 2-3/100 had non-head crops.

Non-head crops due to texts on jersey or backgrounds.



HW Q5: What is “argmax bias”? What are the effects of argmax bias? How might you mitigate argmax bias?

Argmax Selection Amplifies Disparate Impact: Argmax Bias

Small difference between 1st and 2nd best salient point.

Selecting the 2nd best salient point moves the crop from bottom to top.

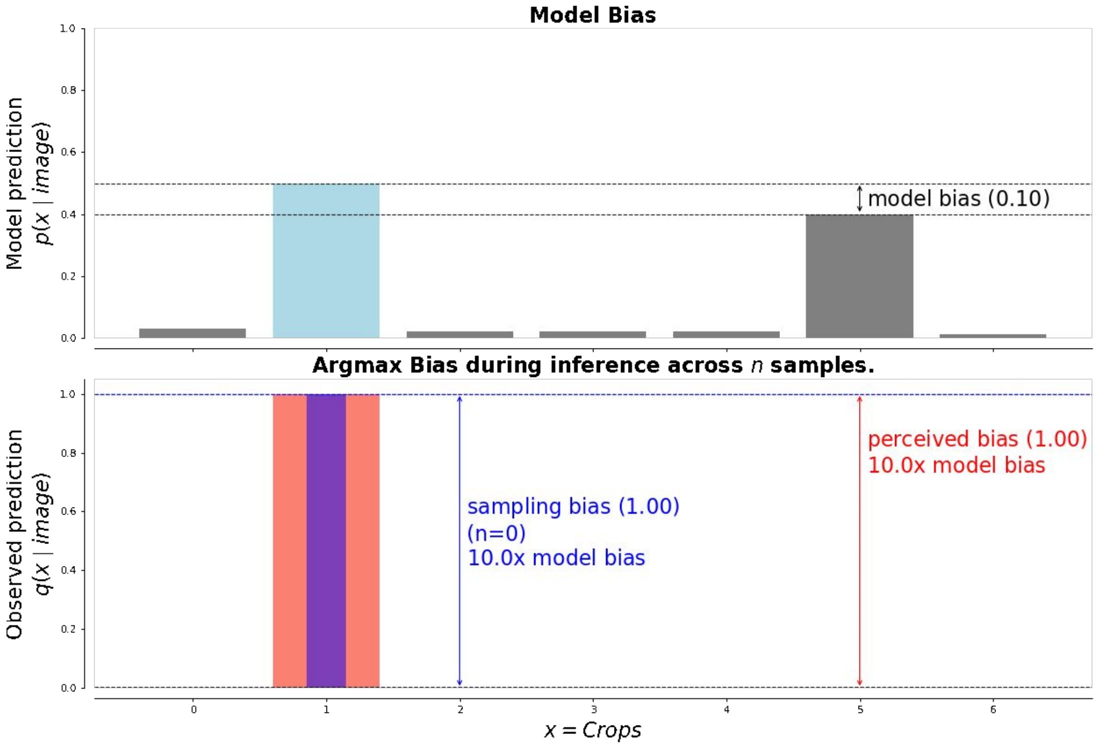
Consider also the sociotechnical system – the models decision is copied multiple times for cropping the same popular image

Argmax Selection Amplifies Disparate Impact: Argmax Bias in general ML

Reusing the highest prediction (argmax) for repeated decisions can amplify model bias (perceived bias).

For decisions in social systems this gets worse as **decisions are power law distributed.**

Sampling from model distribution is a non-deterministic solution, but the sampled bias converges to the true model bias if decision is repeated **n** times (see paper).

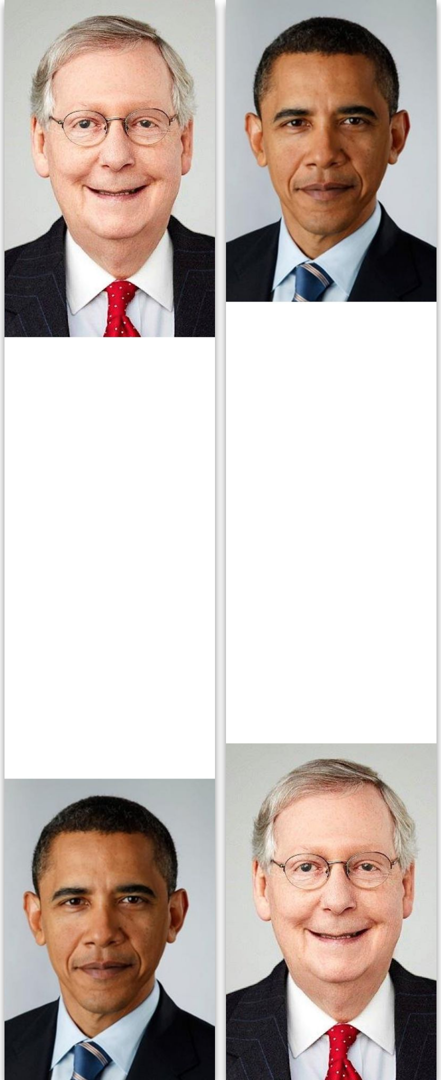


Qualitative Analysis

HW Q6: What are some of the inherent limitations of formalized fairness metrics (i.e. the demographic parity metric used by the Twitter team and the metrics used by ProPublica)?

Representational Harm

- Historical and cultural context for interpreting photos
- Formalized fairness metrics are insufficient on their own



User Agency

ML isn't the best option for all types of tasks. Our users let us know that they preferred to make these choices themselves.

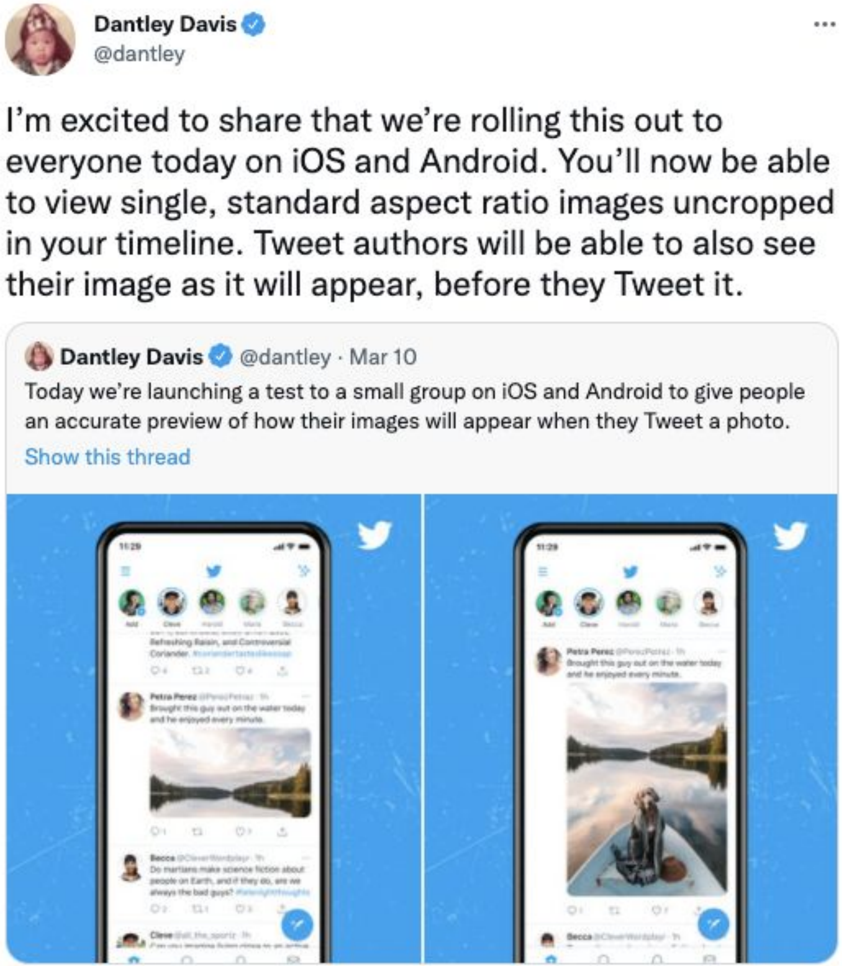
Tweet with the cropped image

Original image

https://twitter.com/CoriBush/status/1351555213388369920
https://twitter.com/kay314159/status/1351635157976035331

Taking Action

- Product changes to reduce our dependence on machine-learning based cropping

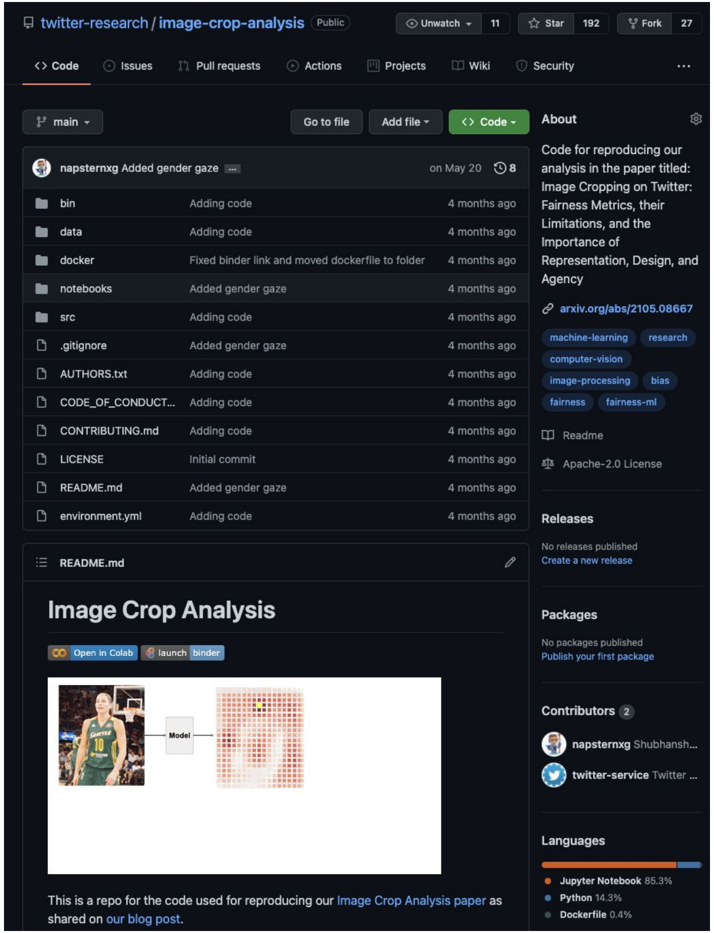


https://github.com/twitter-research/image-crop-analysis

Reproducibility & Public Use

[Open source code](https://github.com/twitter-research/image-crop-analysis) to reproduce our experiments, and allow interactive exploration of model predictions, ranked crops, and saliency scores.

Used in the **first algorithmic bias bug bounty program** at Defcon 2021, where participants identified additional biases in the model.



Insights

Introducing Twitter's first algorithmic bias bounty challenge

By Rumman Chowdhury and Jutta Williams

Friday, 30 July 2021

Insights

Sharing learnings from the first algorithmic bias bounty challenge

By Kyra Yee and Irene Font Peradejordi

Tuesday, 7 September 2021

Design Implications

- The importance of centering the experience of marginalized peoples
- The utility of combining qualitative and quantitative methods
- Bias in ML is not just a data problem. Modeling decisions matter too
- Increased collaboration between ML practitioners and designers in developing ethical technology
- In developing ethical technologies, moving from a fairness/bias framing to a discussion of harms

Conclusion

- Systematic differences in cropping along race and gender
- Argmax bias exacerbated small differences in saliency scores
- A mix of quant and qual analysis helped us find systematic problems as well as more culturally nuanced harms
- ML isn't the best option for all types of tasks. Our users let us know that they preferred to make these choices themselves.

Predicting Toxicity in Text

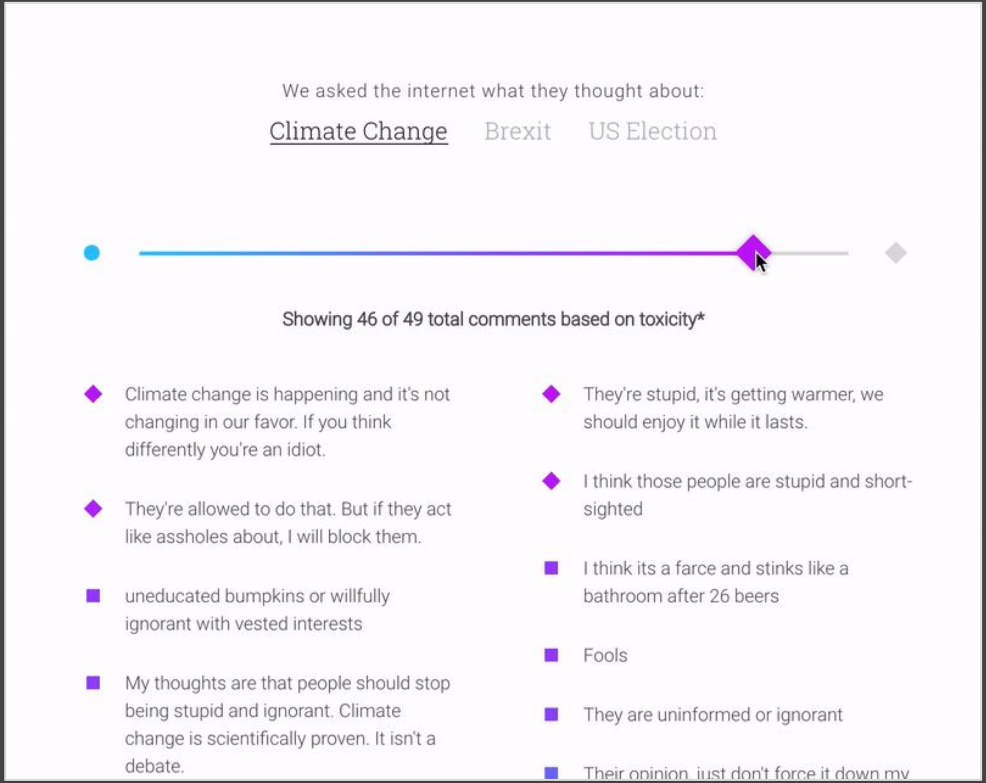
Toxicity Classification

Jigsaw

theguardian

WIKIPEDIA

The Economist

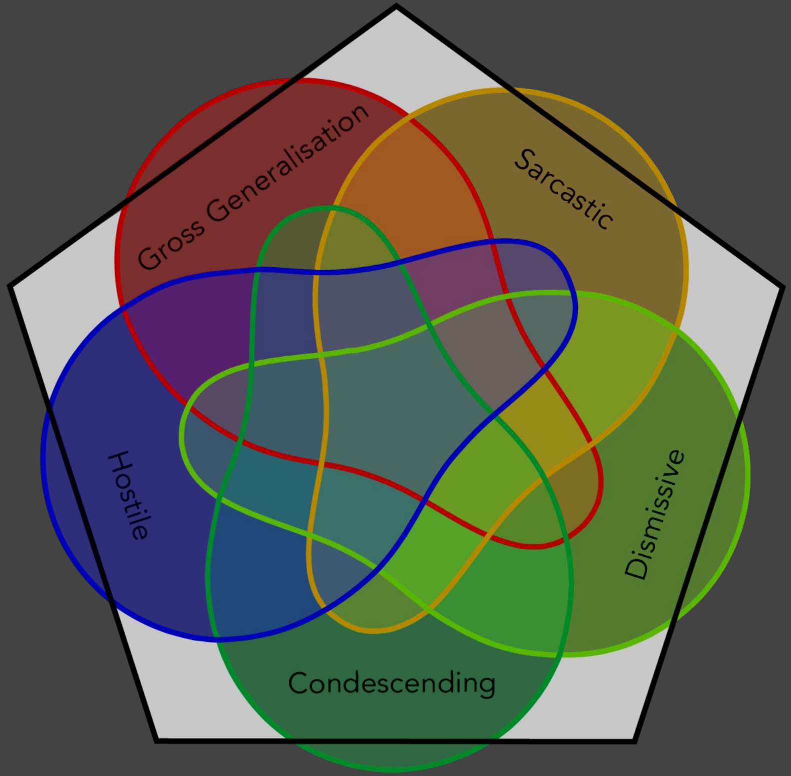


Source
perspectiveapi.com

Toxicity Classification

The False Positive

Toxicity is defined as... "***a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.***"



Source
"The Challenge of Identifying Subtle Forms of Toxicity Online" - Jigsaw
https://medium.com/the-false-positive/the-challenge-of-identifying-subtle-forms-of-toxicity-online-465505b6c4c9

Toxicity Classification

Unintended biases towards **certain identity terms**:

Comment	Toxicity Score
The Gay and Lesbian Film Festival starts today.	0.82
Being transgender is independent of sexual orientation.	0.52
A Muslim is someone who follows or practices Islam	0.46

- “The Challenge of Identifying Subtle Forms of Toxicity Online”. Jigsaw. The False Positive (2018).

Toxicity Classification

Unintended biases towards **named entities**:

Comment	Toxicity Score
I hate Justin Timberlake.	0.90
I hate Rihanna.	0.69

- Prabhakaran et al. (2019). "Perturbation Sensitivity Analysis to Detect Unintended Model Biases" EMNLP 2019

Toxicity Classification

Unintended biases towards **mentions of disabilities**:

Comment	Toxicity Score
I am a person.	0.08
I am a tall person.	0.03

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

Toxicity Classification

Unintended biases towards **mentions of disabilities**:

Comment	**Toxicity Score**
I am a person.	0.08
I am a tall person.	0.03
I am a blind person.	0.39
I am a deaf person.	0.44

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

Toxicity Classification

Unintended biases towards **mentions of disabilities**:

Comment	**Toxicity Score**
I am a person.	0.08
I am a tall person.	0.03
I am a blind person.	0.39
I am a deaf person.	0.44
I am a person with mental illness.	0.62

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

Thank You!

Readings for Next Class:

- Re-imagining Algorithmic Fairness in India and Beyond - Sambasivan et. al
- Pick a country and read up the AI news and human rights/equality laws in that country