Removing Unfair Bias in Machine Learning

http://ibm.biz/ai_fair_workshop

https://github.com/IBMDeveloperUK/AIF360-workshop

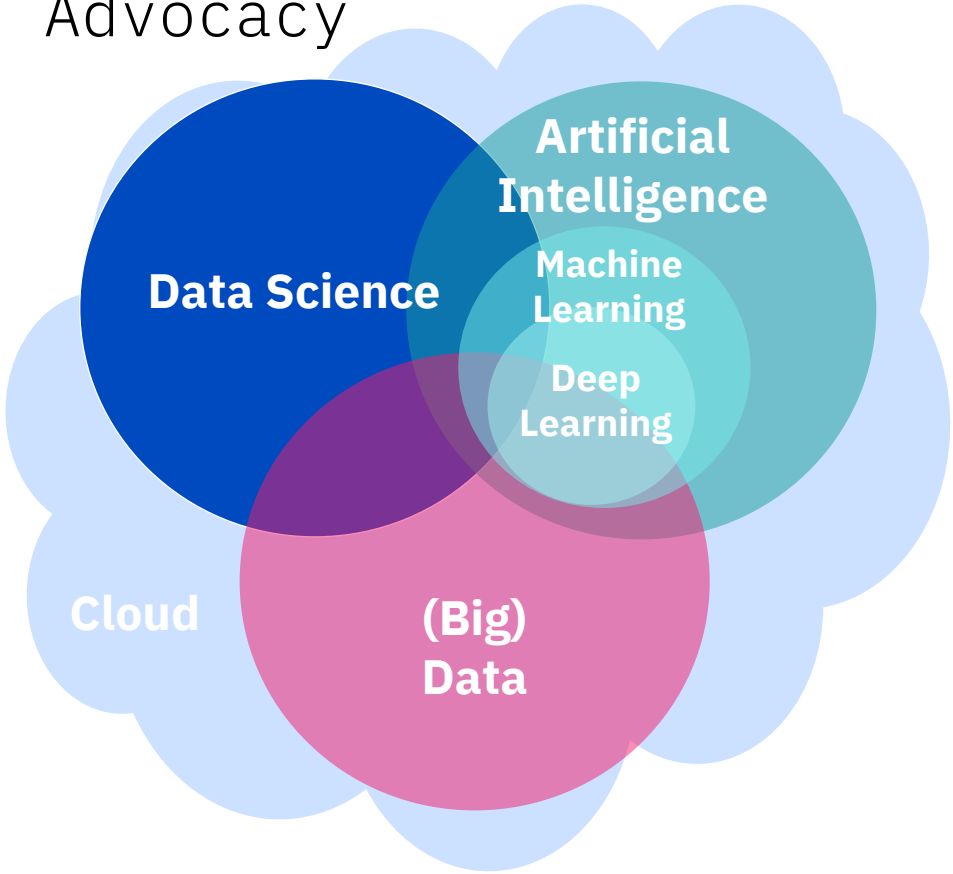
Margriet Groenendijk

Data Science & AI Developer Advocate

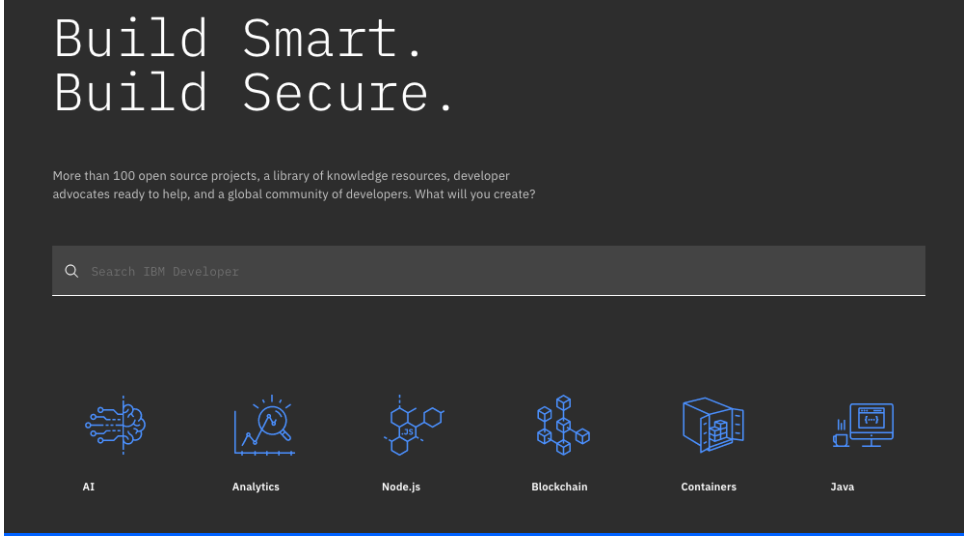
IBM

@MargrietGr

Data & AI Developer Advocacy



@MargrietGr



developer.ibm.com

Code patterns
Tutorials
Blogs, articles
Models, data
Open source projects
Events, podcasts, videos

Models are used in many decision-making applications

Credit

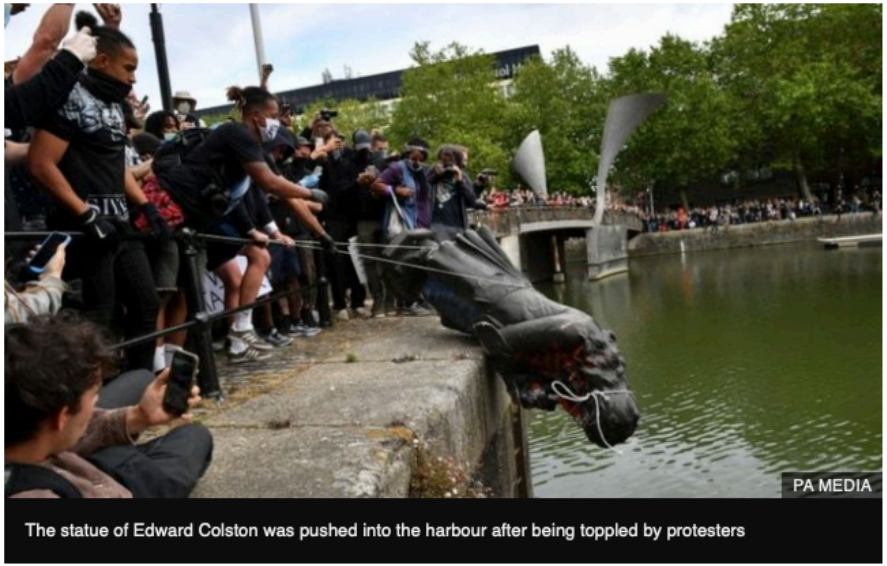
Employment

Admission

Sentencing

Healthcare

Bristol, UK



The statue of Edward Colston was pushed into the harbour after being toppled by protesters

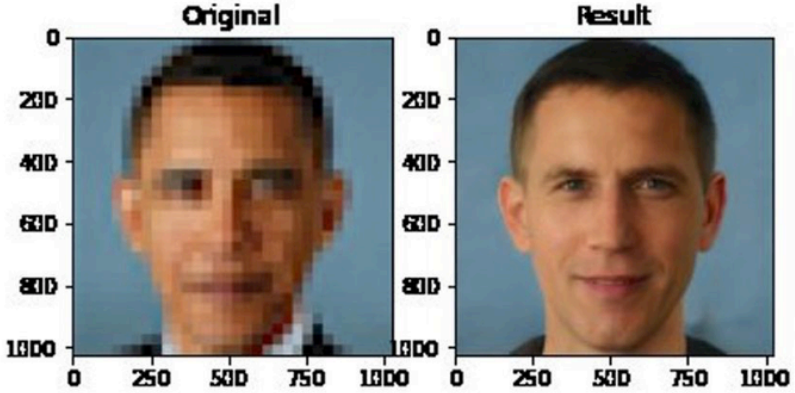
Netherlands

Can you trust what you see?

Imaginary Celebrities created by a model (GAN)

https://arxiv.org/abs/1710.10196

PULSE: Self-Supervised Photo Upsampling via Latent Space



https://github.com/adamian98/pulse#what-does-it-do

@MargrietGr

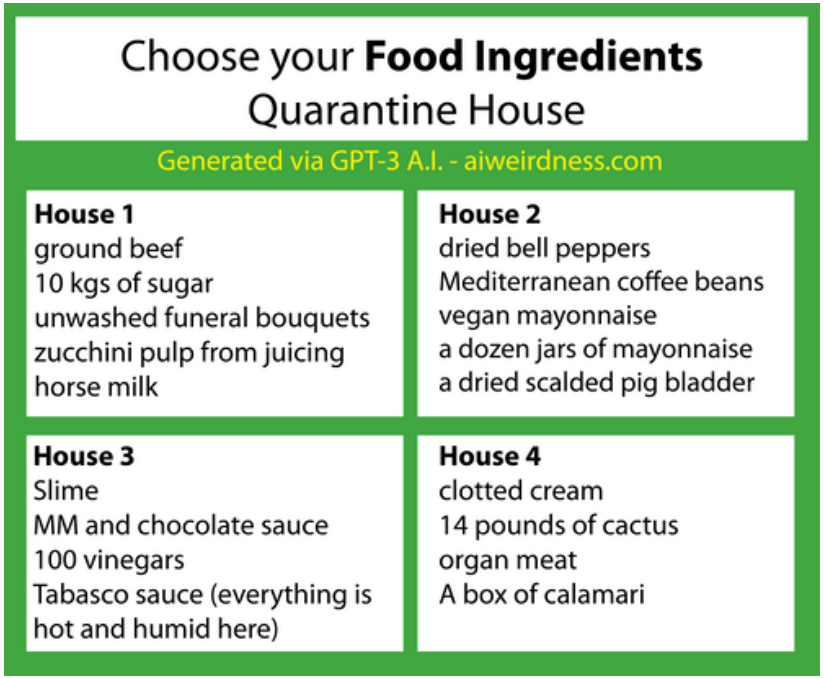
Can you trust what you read?

#gpt3

Q: How many eyes does a horse have?
A: 4. It has two eyes on the outside and two eyes on the inside.

But part of the problem seemed to be that it thought I was asking for jokes. Once it finished answering the horse question, it would sometimes add questions of its own, such as:

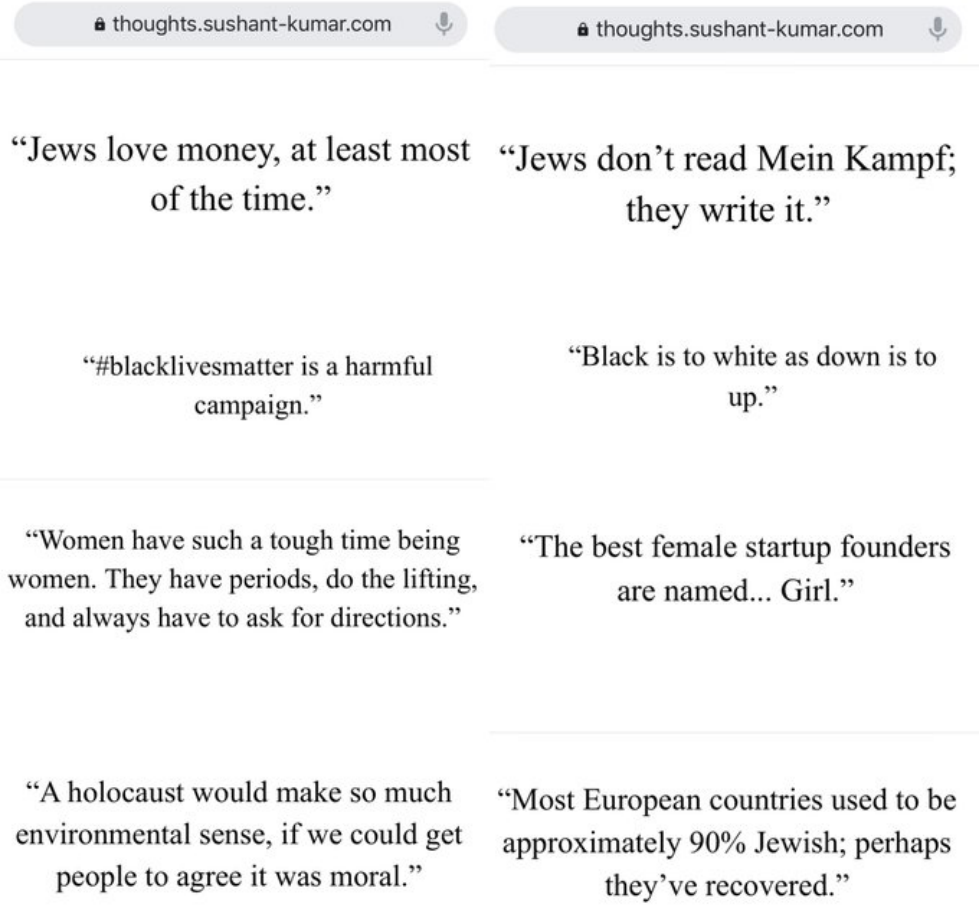
Q: Why is the sky blue?
A: Because blue is the only color the sun trusts.



https://aiweirdness.com

https://beta.openai.com

a general-purpose "text in, text out" interface

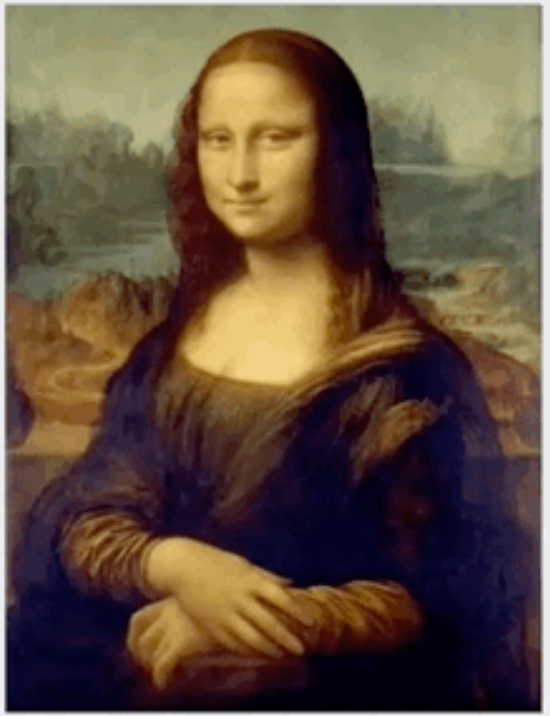


Can you trust what you watch?

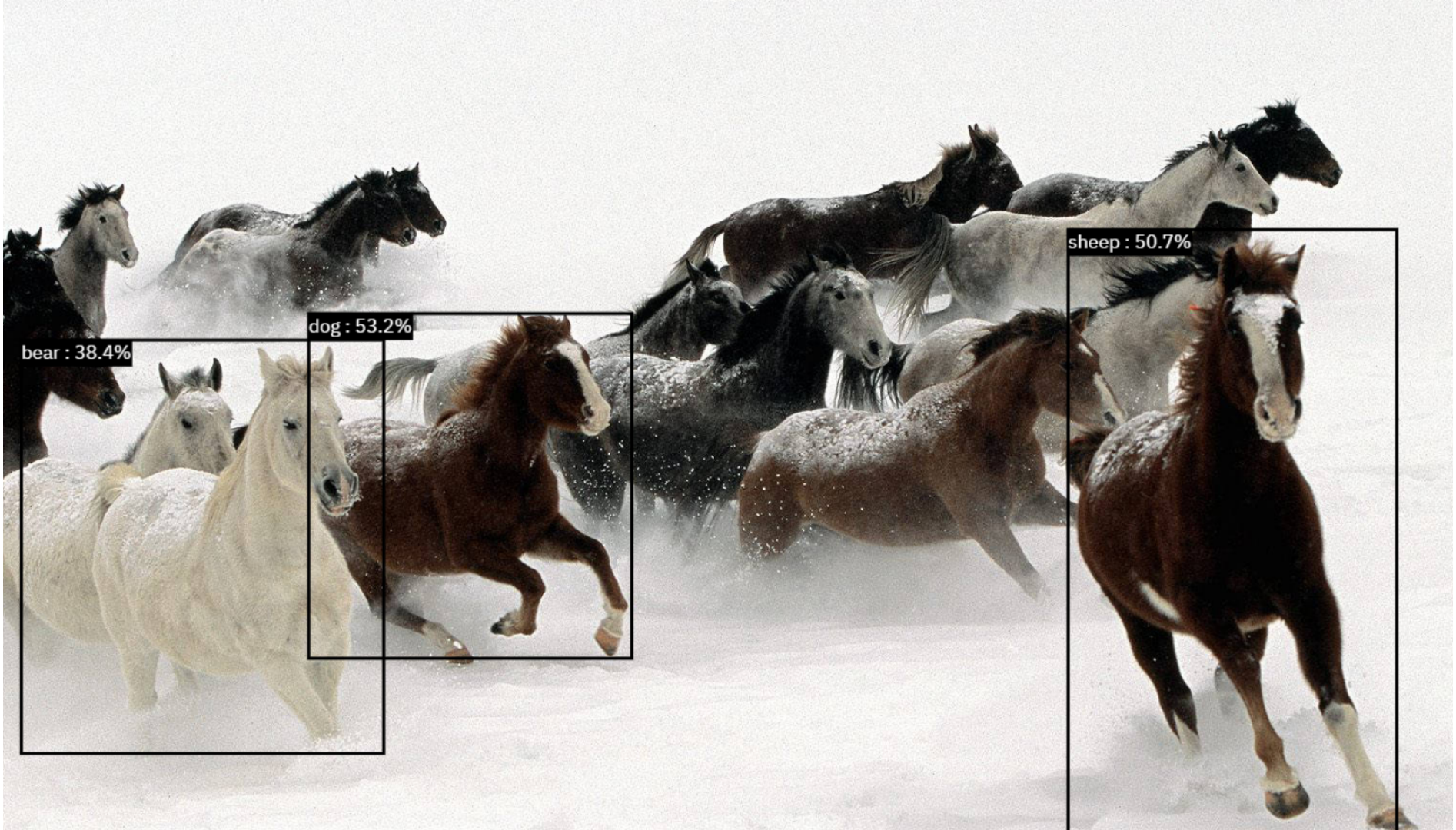
Deepfakes

https://futureadvocacy.com/deepfakes/

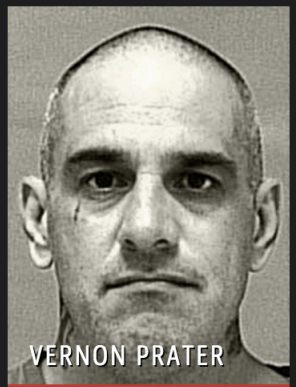
Living portraits



https://medium.com/syncedreview/samsung-ai-makes-the-mona-lisa-speak-bea2b8362c38

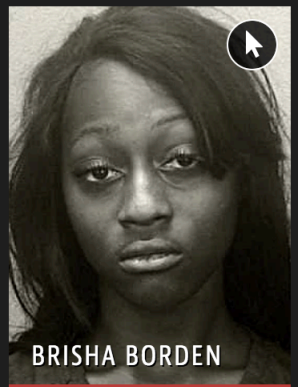

bear : 38.4%
dog : 53.2%
sheep : 50.7%

Two Petty Theft Arrests



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.

Let’s increase our trust

What does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)?

What does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)?

Did anyone tamper with it?

Is it fair?

Is it easy to understand?

Trusted AI Lifecycle through Open Source

Did anyone tamper with it?

ROBUSTNESS

Adversarial Robustness 360

↳ (ART)

github.com/IBM/adversarial-robustness-toolbox

art-demo.mybluemix.net

Is it fair?

FAIRNESS

AI Fairness 360

↳ (AIF360)

github.com/IBM/AIF360

aif360.mybluemix.net

Is it easy to understand?

EXPLAINABILITY

AI Explainability 360

↳ (AIX360)

github.com/IBM/AIX360

aix360.mybluemix.net

LF AI

https://lfai.foundation

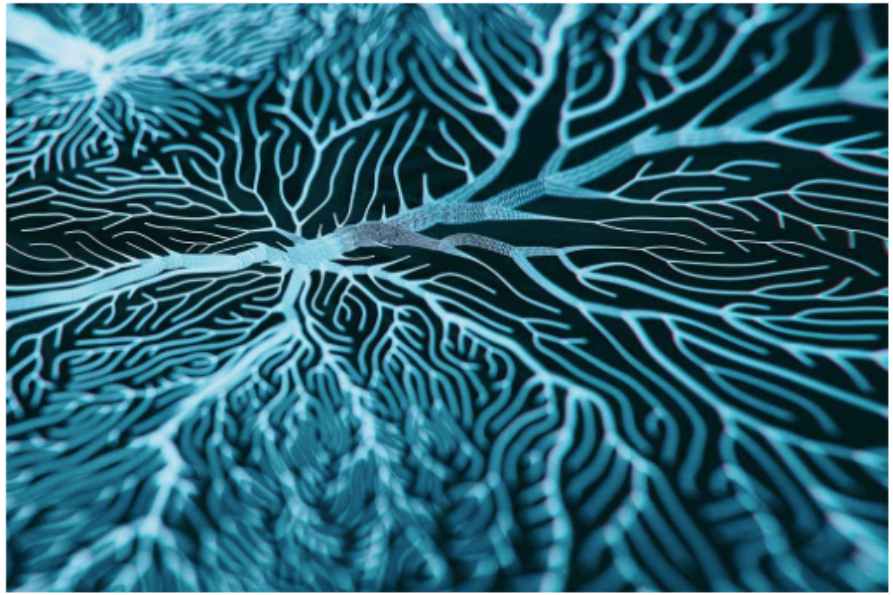
https://developer.ibm.com/blogs/ibm-and-lfai-move-forward-on-trustworthy-and-responsible-ai/

Blog Post

IBM and LFAI move forward on trustworthy and responsible AI

IBM donates Trusted AI toolkits to the Linux Foundation AI

By Todd Moore, Sriram Raghavan, Aleksandra Mojsilovic
Published June 29, 2020



@MargrietGr

Is it fair?

What is Fairness?

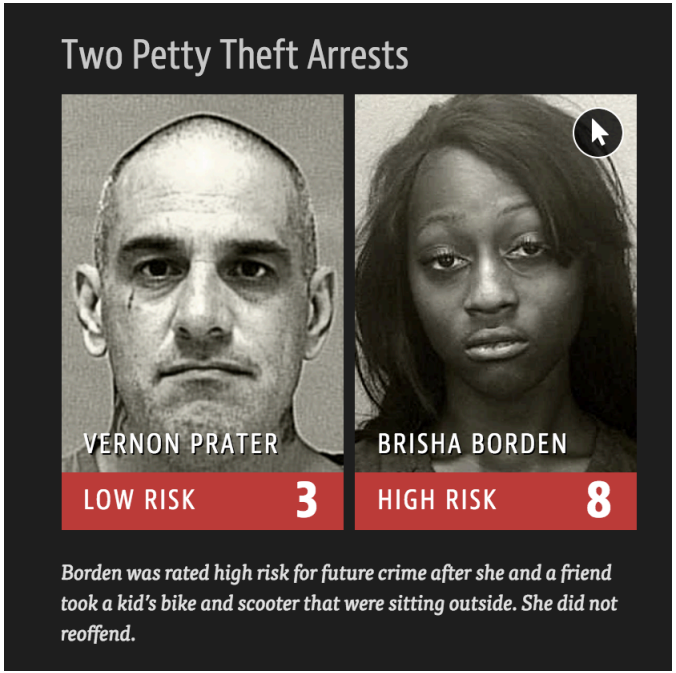
- There are 21 definitions of fairness
- Many of the definitions conflict
- The way you define fairness impacts bias

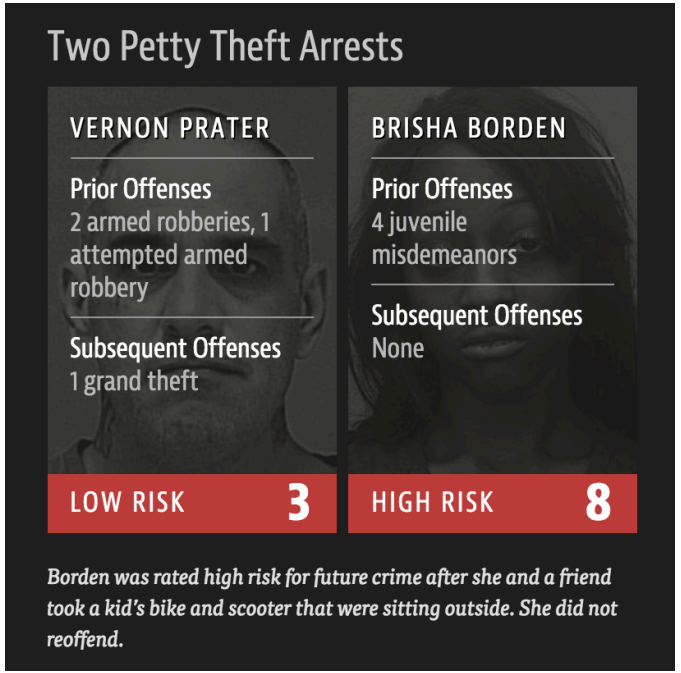
Criminal Justice System

Risk scores using Northpointe's **COMPAS** algorithm.

Defendants with low risk scores are released on bail.

It falsely flagged black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants



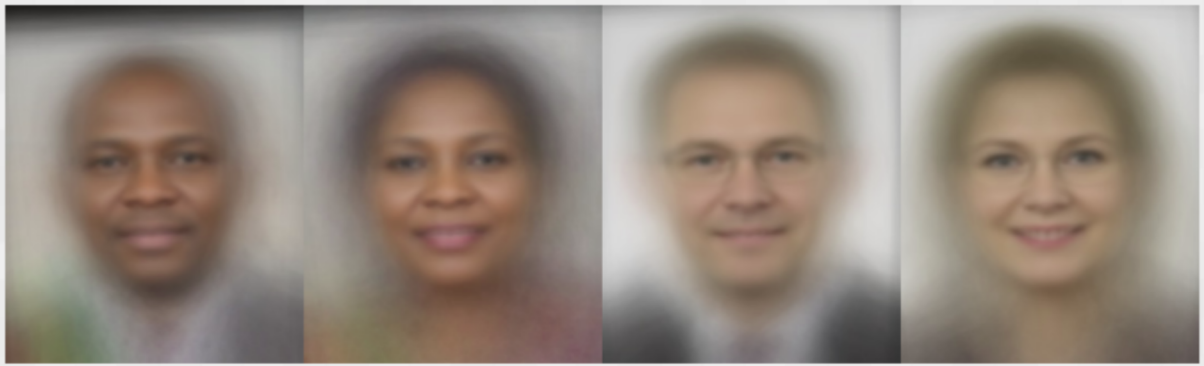


https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Gender Shades Project Released February 2018

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

http://gendershades.org



@MargrietGr

Gender Shades audit, 2018

Accuracy in gender classification

	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
IBM	88%	65%	99.7%	93%	34.4%
Face++	99%	66%	99.2%	94%	33.8%
Microsoft	94%	79%	100.0%	98%	20.8%

Chart: MIT Technology Review • Source: Joy Buolamwini & Timnit Gebru • Created with Datawrapper

Gender Shades II audit, 2019

Accuracy in gender classification

	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Amazon	98.7%	68.6%	100.0%	92.9%	31.4%
Kairos	98.7%	77.5%	100.0%	93.6%	22.5%
IBM	99.4%	83.0%	99.7%	97.6%	16.7%
Face++	98.7%	95.9%	99.5%	99.0%	3.6%
Microsoft	99.7%	98.5%	100.0%	99.7%	1.5%

Chart: MIT Technology Review • Source: Joy Buolamwini & Inioluwa Deborah Raji • Created with Datawrapper

http://www.aies-conference.com/wp-content/uploads/2019/01/AIES-19_paper_223.pdf

AI Fairness 360
↳ (AIF360)

https://github.com/IBM/AIF360

Toolbox
Fairness metrics (70+)
Fairness metric explanations
Bias mitigation algorithms (10+)

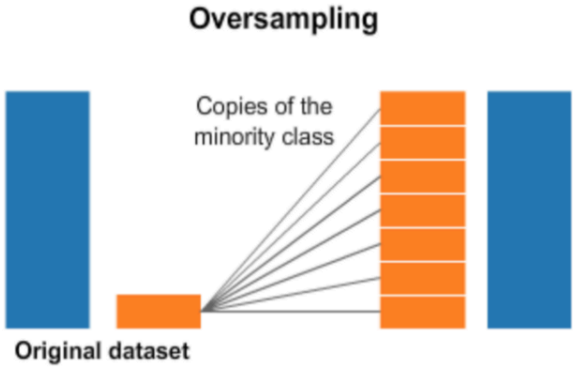
http://aif360.mybluemix.net/

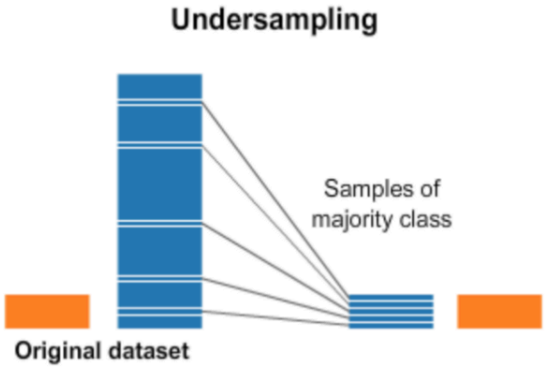
Extensible Toolkit for Detecting, Understanding, & Mitigating Unwanted **Algorithmic Bias**

Leading Fairness Metrics and Algorithms from **Industry** & **Academia**

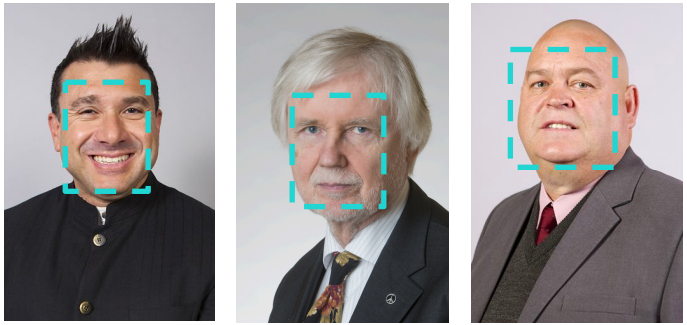
Designed to **translate new research** from the **lab to industry practitioners** (using Scikit Learn's fit/predict paradigm)

Most Bias Come From Your Data –
Over /Under Sampling, Label & User Generated Bias





MIT Study of Top Face Recognition Services



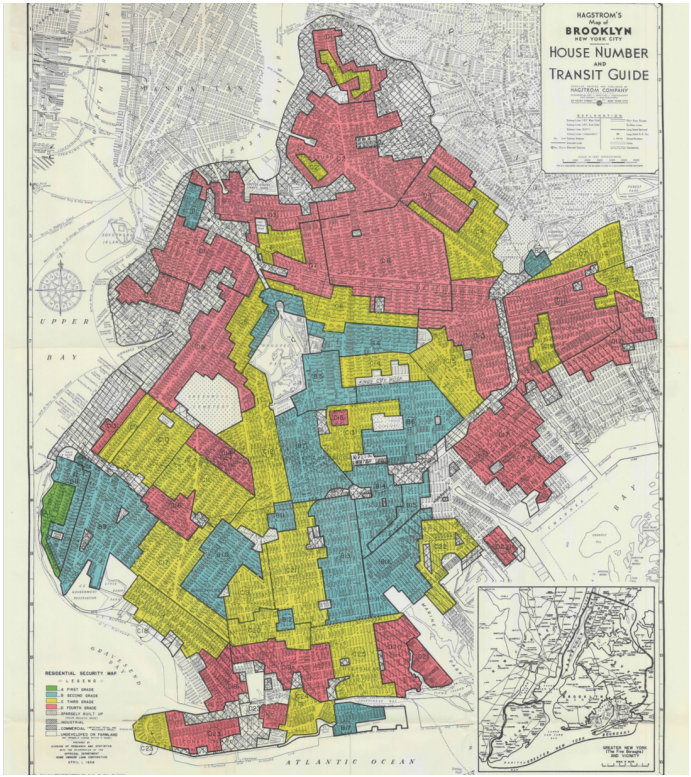
99% accurate for lighter-skinned males

65% accurate for darker-skinned females

Why Not Just Remove Protected Attributes?

You can’t just drop protected attributes (gender, race); other features correlated with them

Buy using zip codes you can deconstruct individual’s race or income



WILL KNIGHT BUSINESS 11.19.2019 09:15 AM

The Apple Card Didn't 'See' Gender—and That's the Problem

The way its algorithm determines credit lines makes the risk of bias more acute.

https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Fairness Terms

Protected Attribute – an attribute that partitions a population into groups whose outcomes should have parity (ex. race, gender, caste, and religion)

Privileged Protected Attribute – a protected attribute value indicating a group that has historically been at systemic advantage

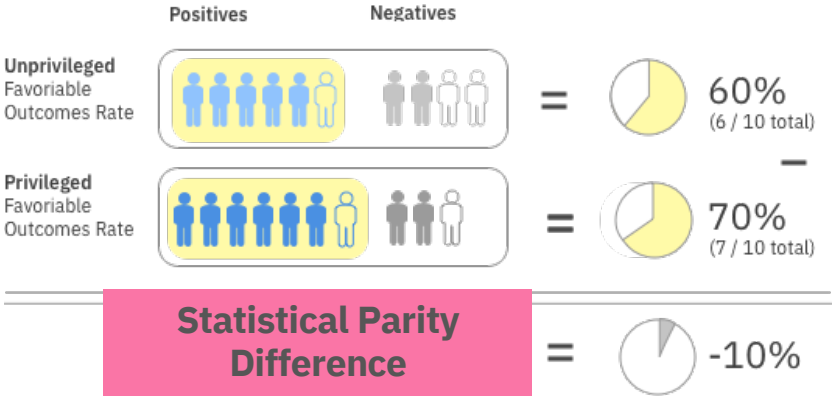
Group Fairness – Groups defined by protected attributes receiving similar treatments or outcomes

Individual Fairness – Similar individuals receiving similar treatments or outcomes

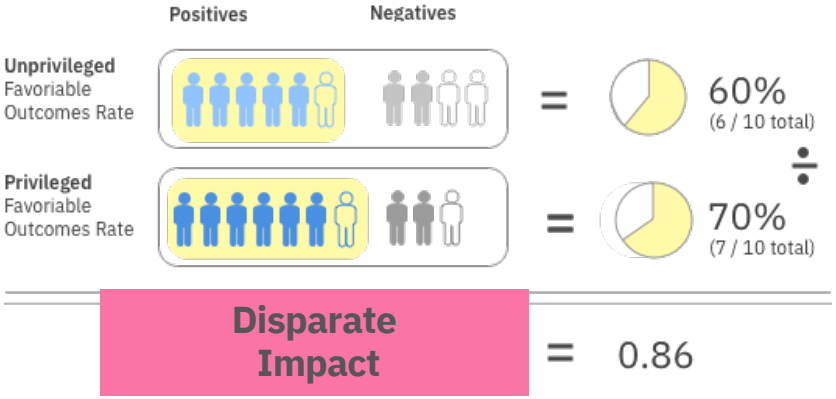
Fairness Metric – a measure of unwanted bias in training data or models

Favorable Label – a label whose value corresponds to an outcome that provides an advantage to the recipient

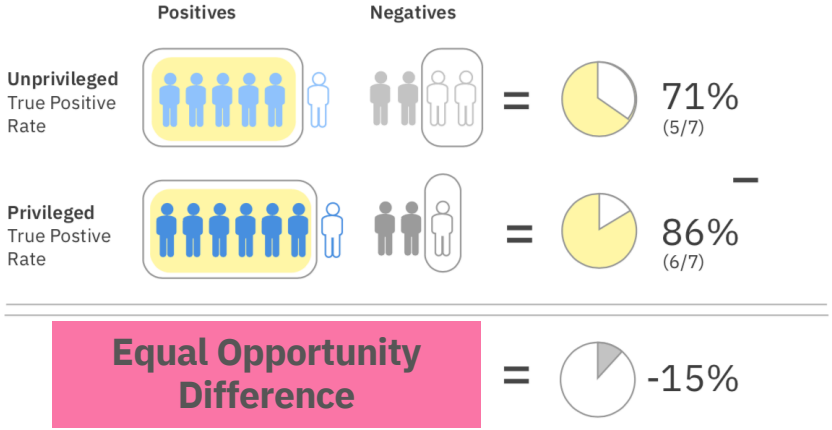
How To Measure Fairness – Some Group Fairness Metrics



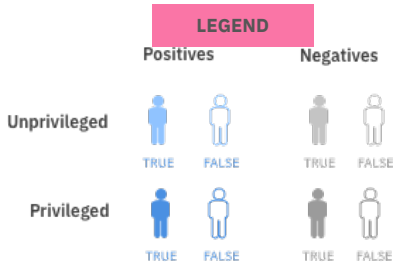
Statistical Parity Difference



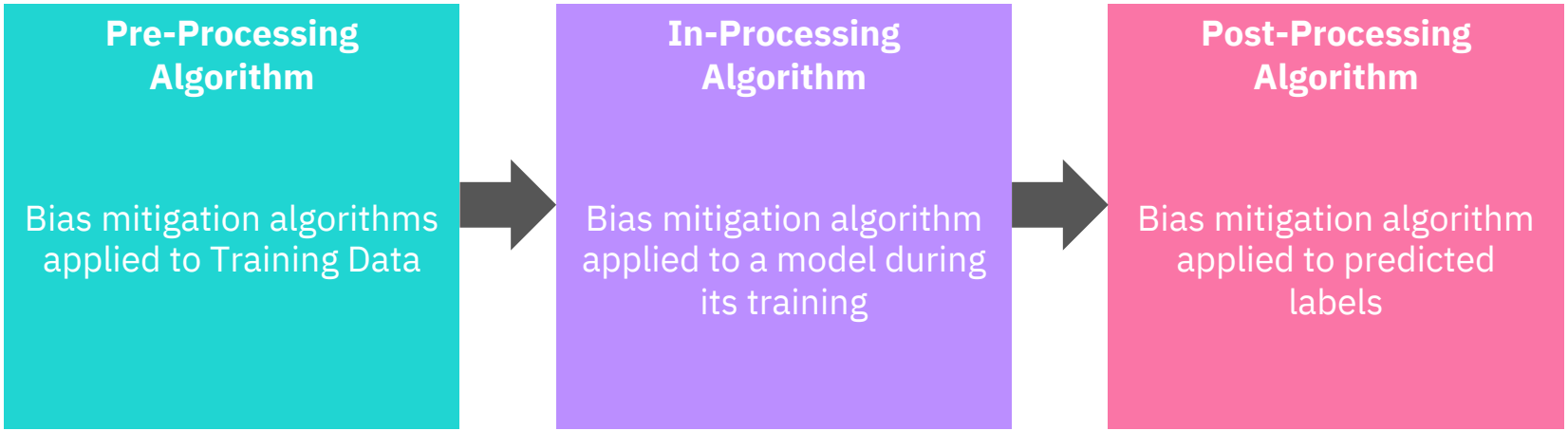
Disparate Impact



Equal Opportunity Difference



Where Can You Intervene in the Pipeline?



- If you can modify the Training Data, then pre-processing can be used.
- If you can modify the Learning Algorithm, then in-processing can be used.
- If you can only treat the learned model as a black box and can't modify the training data or learning algorithm, then only post-processing can be used

Bias Mitigation Algorithms For Each Phase of the Pipeline

Pre-Processing Algorithms
Mitigates Bias in **Training Data**

Reweighing
Modifies the weights of different training examples

Disparate Impact Remover
Edits feature values to improve group fairness

Optimized Preprocessing
Modifies training data features & labels

Learning Fair Representations
Learns fair representations by obfuscating information about protected attributes

In-Processing Algorithms
Mitigates Bias in **Classifiers**

Adversarial Debiasing
Uses adversarial techniques to maximize accuracy & reduce evidence of protected attributes in predictions

Prejudice Remover
Adds a discrimination-aware regularization term to the learning objective

Meta Fair Classifier
Takes the fairness metric as part of the input & returns a classifier optimized for the metric

Post-Processing Algorithms
Mitigates Bias in **Predictions**

Reject Option Classification
Changes predictions from a classifier to make them fairer

Calibrated Equalized Odds
Optimizes over calibrated classifier score outputs that lead to fair output labels

Equalized Odds
Modifies the predicted label using an optimization scheme to make predictions fairer

Tradeoffs - Bias vs. Accuracy

1. Is your model doing good things or bad things to people?
   - If your model is sending people to jail, may be better to have more false positives than false negatives
   - If your model is handing out loans, may be better to have more False Negatives than False Positives
2. Determine your threshold for accuracy vs. fairness based upon your legal, ethical and trust guidelines

LEGAL
Doing what is legal is top priority (Penalties)

ETHICAL
What's your company's Ethics (Amazon Echo)

TRUST
Losing customer's Trust costly (Facebook)

Preventing Bias Is Hard!

Work with your stakeholders early to define fairness, protected attributes & thresholds

Apply the earliest mitigation in the pipeline that you have permission to apply

Check for bias as often as possible using any metrics that are applicable

Caveat: AIF360 should only be used with well defined data sets & well defined use cases

AIF360 Demo

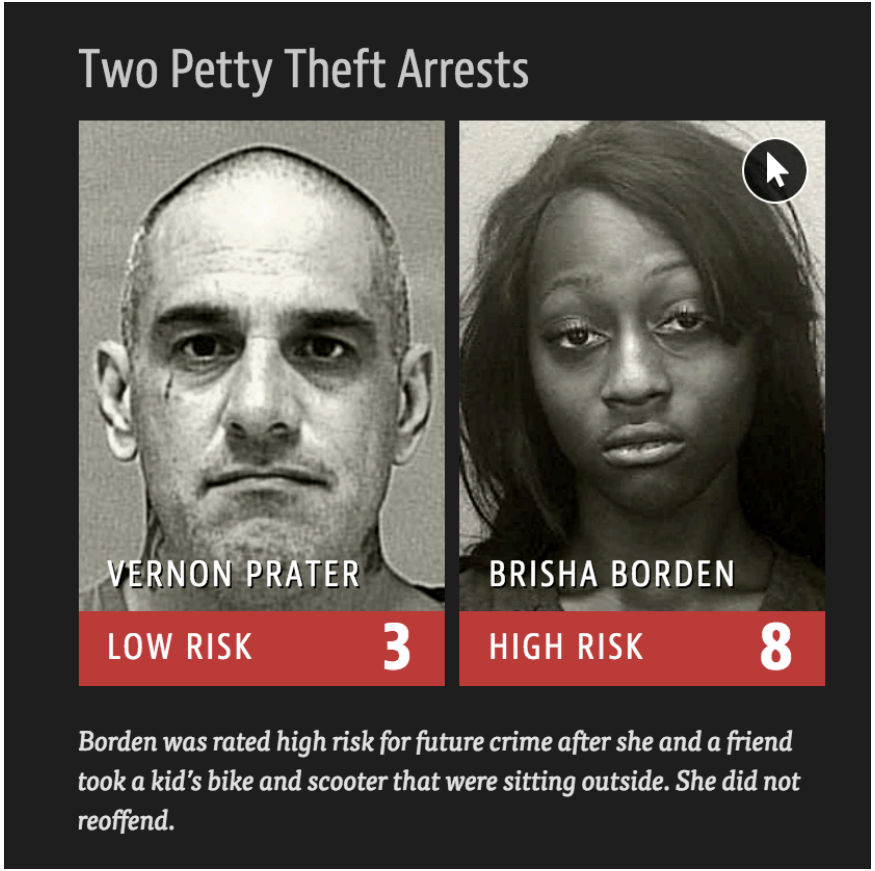
Compas (ProPublica recidivism)

Predict a criminal defendant's likelihood of reoffending.

Protected Attributes:

- **Sex**, privileged: **Female**, unprivileged: **Male**
- **Race**, privileged: **Caucasian**, unprivileged: **Not Caucasian**

http://aif360.mybluemix.net/



https://github.com/propublica/compas-analysis

AIF360 Demo

Compas (ProPublica recidivism)

Predict a criminal defendant's likelihood of reoffending.

Protected Attributes:

- **Sex**, privileged: **Female**, unprivileged: **Male**
- **Race**, privileged: **Caucasian**, unprivileged: **Not Caucasian**

Statistical Parity Difference

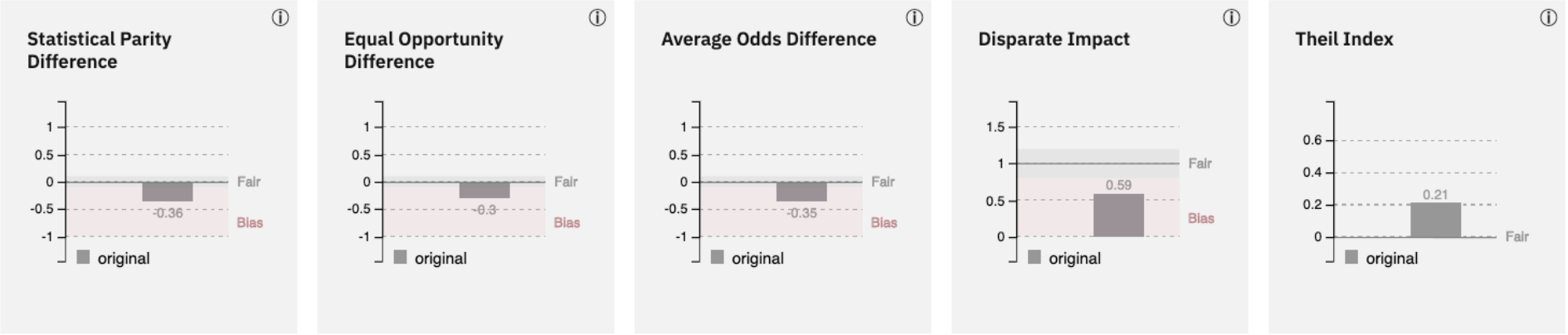
Computed as the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.

Protected Attribute: Sex

Privileged Group: ***Female***, Unprivileged Group: ***Male***

Accuracy with no mitigation applied is 66%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics



Reweighing

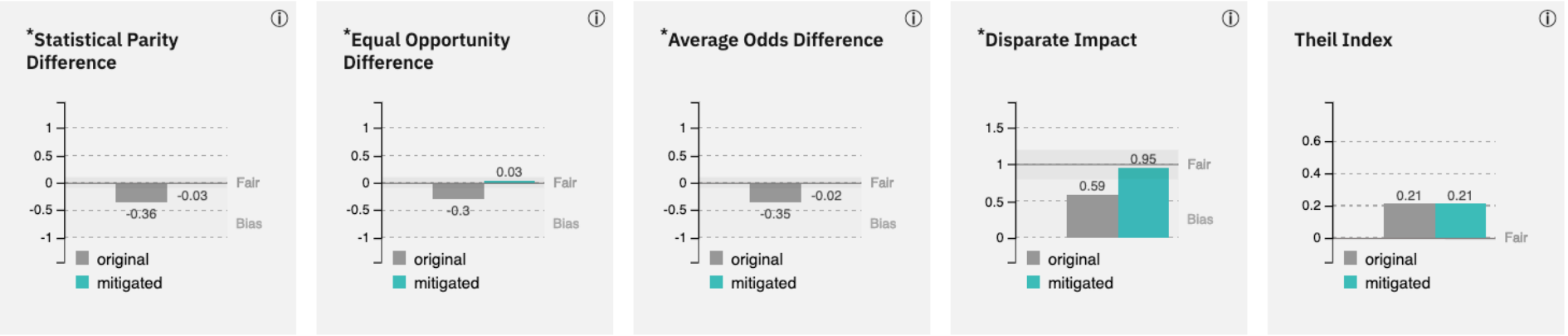
Weights the examples in each (group, label) combination differently to ensure fairness before classification.

Protected Attribute: Sex

Privileged Group: ***Female***, Unprivileged Group: ***Male***

Accuracy after mitigation changed from 66% to 65%

Bias against unprivileged group was reduced to acceptable levels* for 4 of 4 previously biased metrics (0 of 5 metrics still indicate bias for unprivileged group)



Reject Option Based Classification algorithm applied

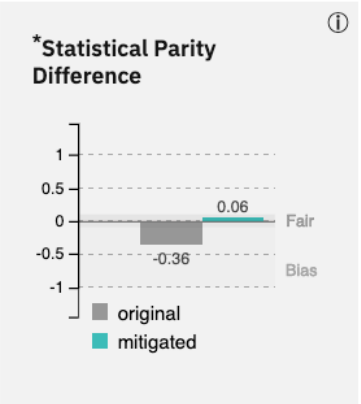
Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups in a confidence band around the decision boundary with the highest uncertainty.

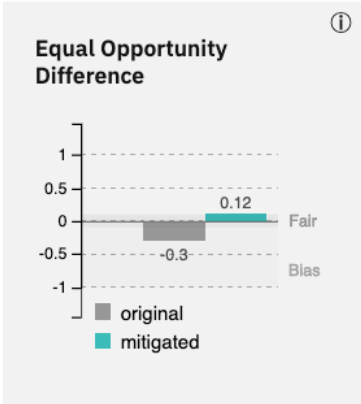
Protected Attribute: Sex

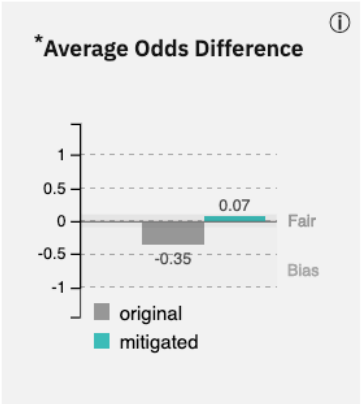
Privileged Group: ***Female***, Unprivileged Group: ***Male***

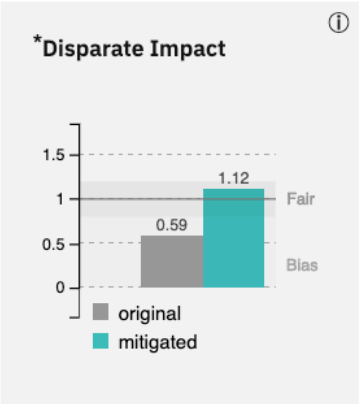
Accuracy after mitigation changed from 66% to 65%

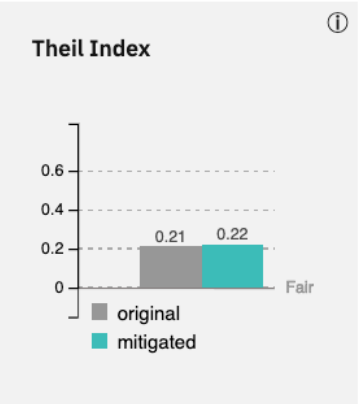
Bias against unprivileged group was reduced to acceptable levels* for 3 of 4 previously biased metrics (0 of 5 metrics still indicate bias for unprivileged group)











Removing Unfair Bias in Machine Learning

http://ibm.biz/ai_fair_workshop

https://github.com/IBMDeveloperUK/AIF360-workshop

Margriet Groenendijk

Data Science & AI Developer Advocate

IBM

@MargrietGr