

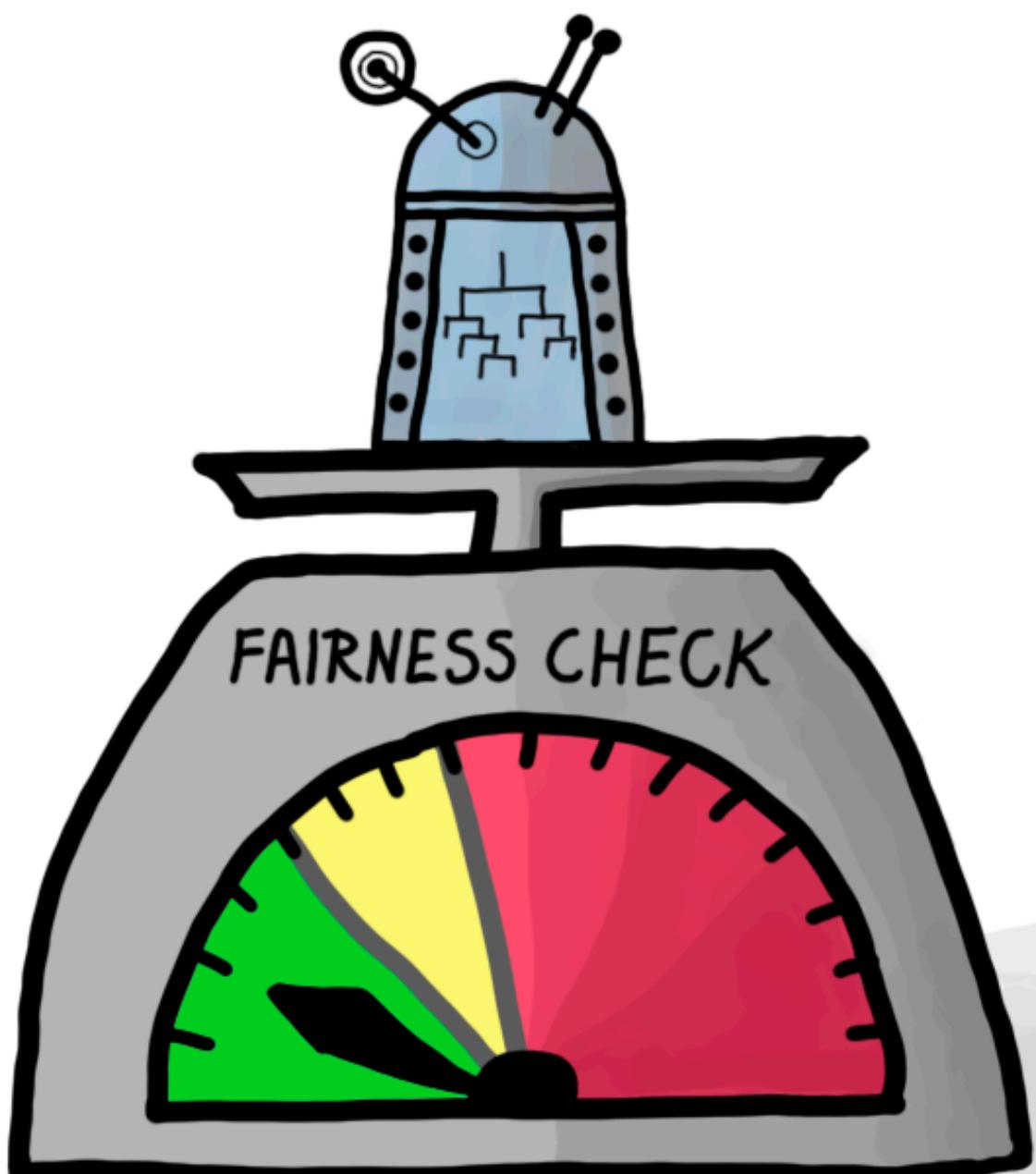
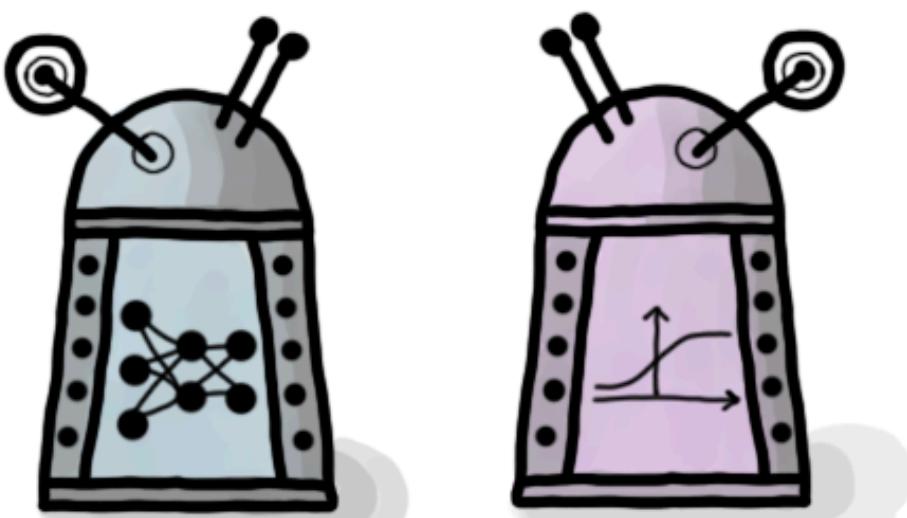
# Explaining and Checking Fairness for Predictive Models

Przemysław Biecek

3rd Workshop  
eXplaining Knowledge Discovery in Data Mining 2021

Materials: <https://tinyurl.com/xkdd-fairness>

Slack: <https://tinyurl.com/xkdd21>



Do algorithms  
discriminate?


**MACHINE BIAS**

# Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement

New research and Facebook's own ad archive show that the company's new system to ensure diverse audiences for housing and employment ads has many of the same problems as its predecessor.

by Ava Kofman and Ariana Tobin, Dec. 13, 2019, 5 a.m. EST

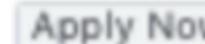
 Inactive  
Nov 4, 2019 - Nov 14, 2019  
ID: 991590397853985   

 About social issues, elections or politics

 Dolese Bros. Co.  
Sponsored • Paid for by Dolese Bros. Co.

A local career to keep you close to home. We're hiring CDL drivers! #DoleseDelivers #CareerOpportunities



Now Hiring CDL Drivers  
Join our team today!  
[WWW.DOLESE.COM](http://WWW.DOLESE.COM) 

[See Ad Details](#)



Left: A Facebook ad for Dolese Bros. Co. Right: Facebook's chart shows that 87 percent of the people who saw the ad were men.

BUSINESS NEWS

OCTOBER 10, 2018 / 5:12 AM / A MONTH AGO

## Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN

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

The group created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes. The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes, the people said.

Instead, the technology favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured," one person said.

**Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist**

THE VERGE

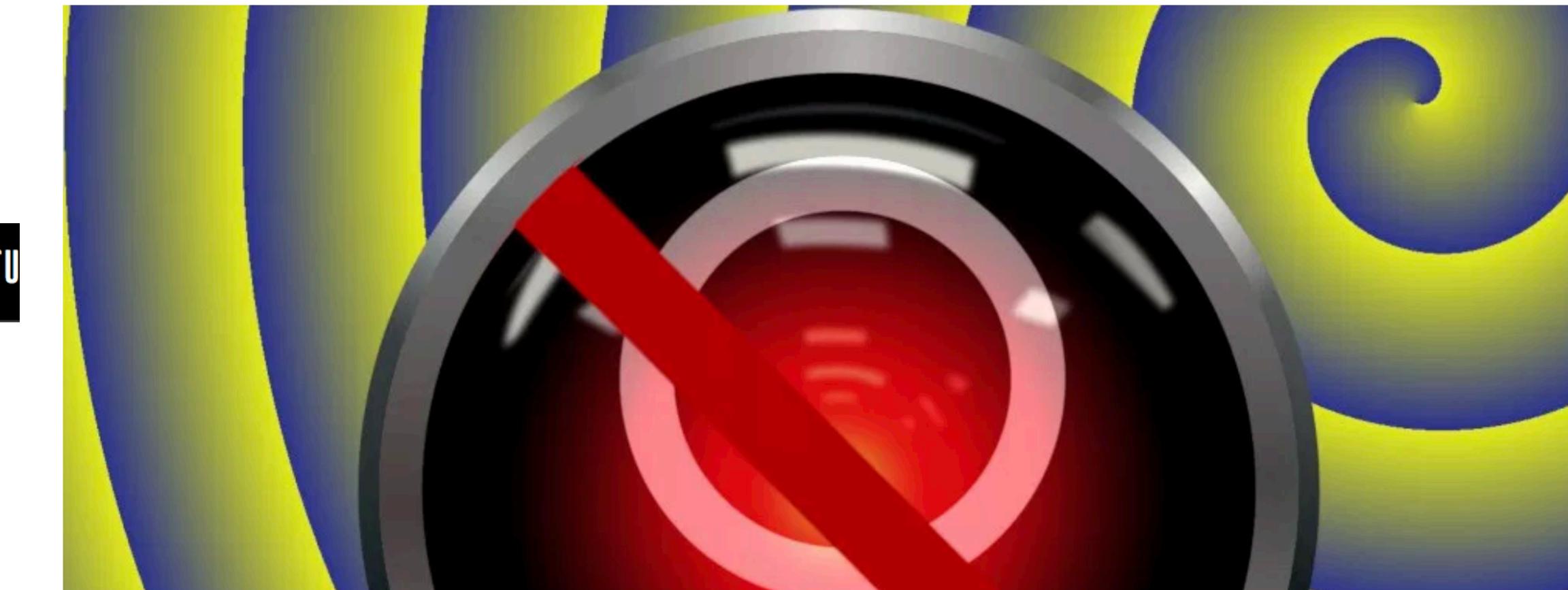
TECH ▾ SCIENCE ▾ CULTURE

TECH ▾ AMAZON ▾ ARTIFICIAL INTELLIGENCE

## Amazon reportedly scraps internal AI recruiting tool that was biased against women

21

*The secret program penalized applications that contained the word "women's"*

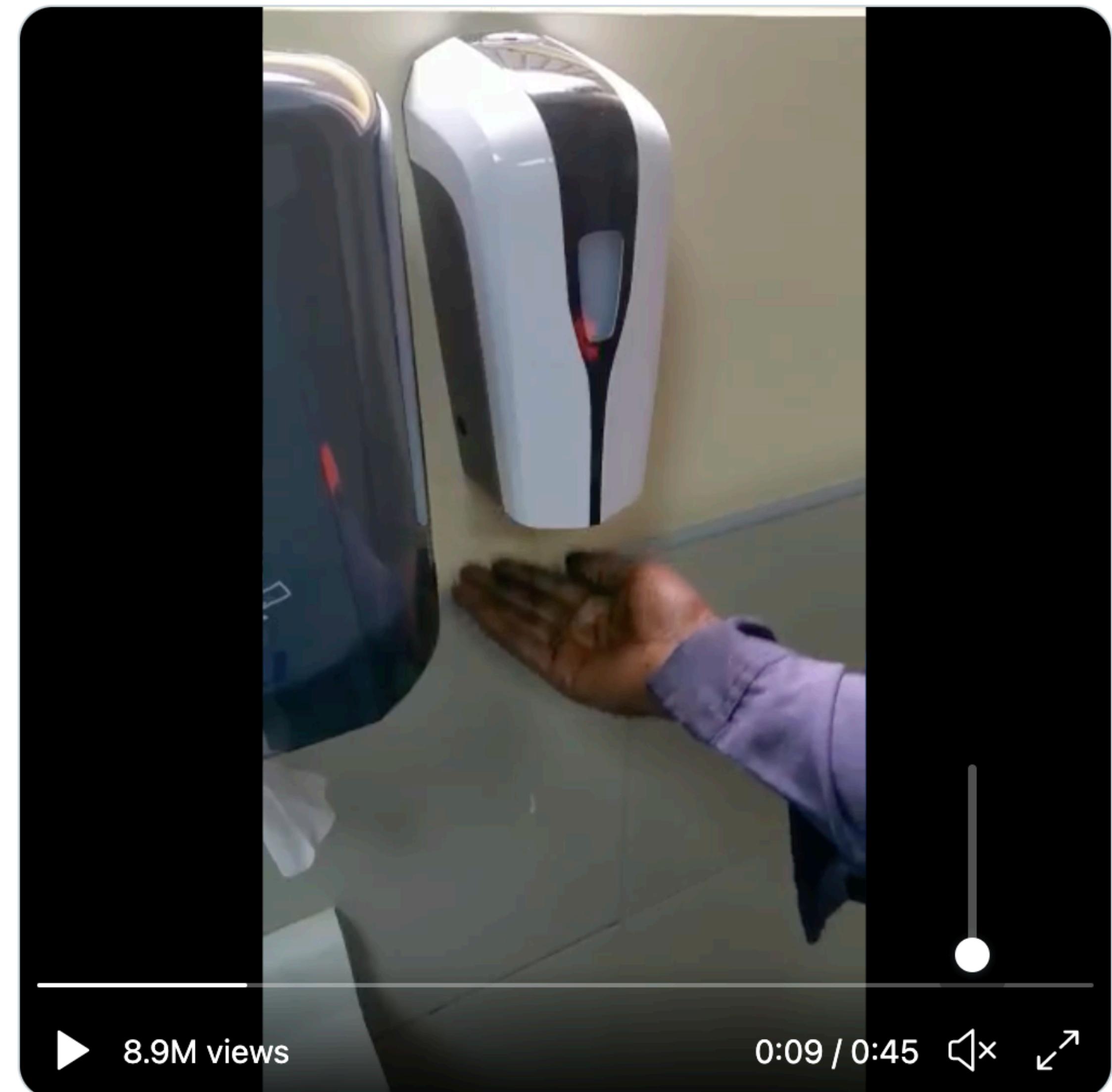


# Racist Soap Dispenser



**Chukwuemeka Afigbo @nke\_ise** · Aug 16, 2017

If you have ever had a problem grasping the importance of diversity in tech and its impact on society, watch this video



2.7K

166.4K

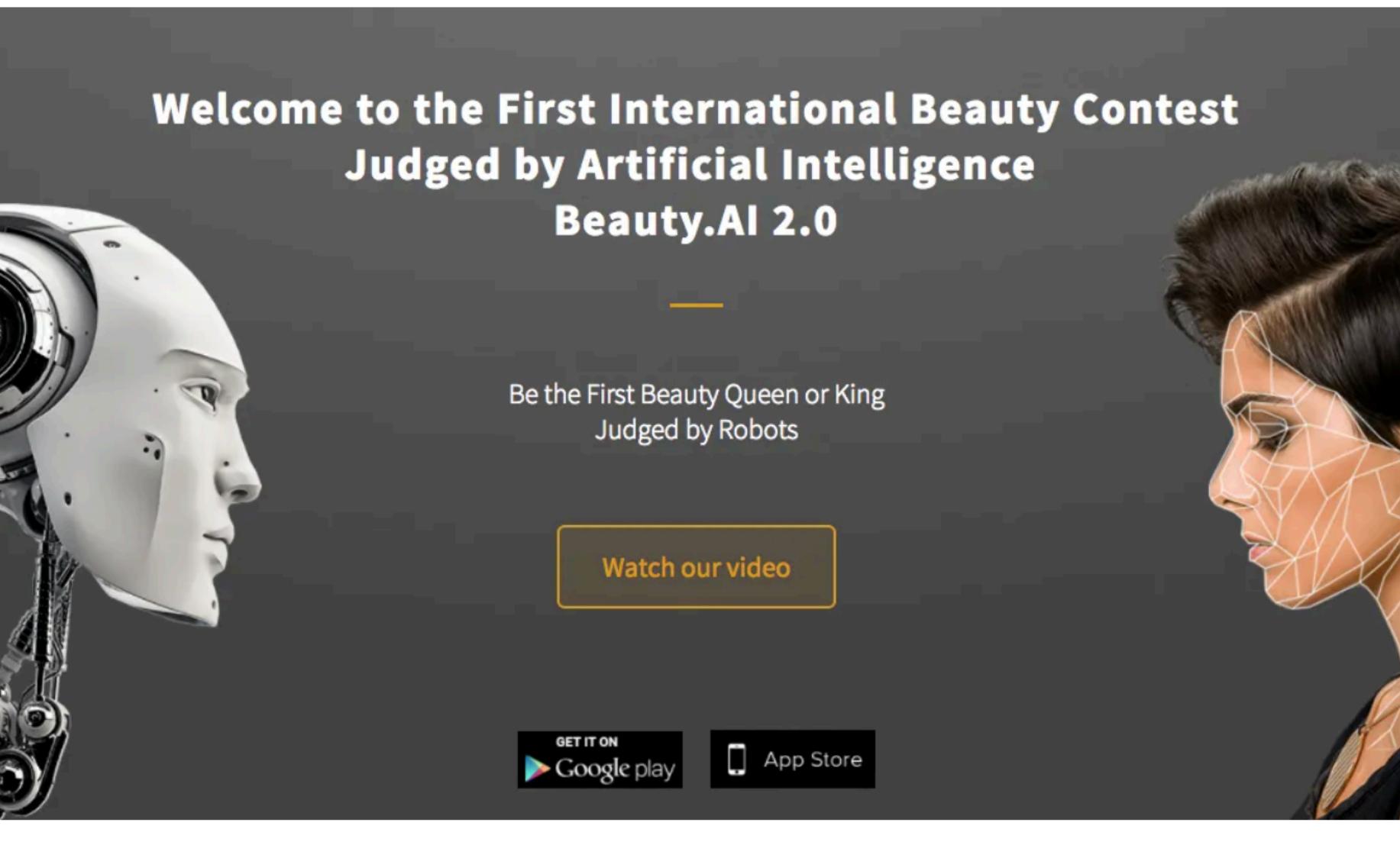
211.3K



[https://twitter.com/nke\\_ise/status/897756900753891328](https://twitter.com/nke_ise/status/897756900753891328)

# Beauty.AI's 'robot beauty contest' is back – and this time it promises not to be racist

The Beauty.AI robot beauty contest discriminated against dark skin. Now it's being relaunched - along with a 'Diversity AI think tank'



*Update 02.03.2017 12:19: Since this article was published, the material on the Diversity.AI website has been taken down and replaced with a single message: "The initiative has not yet been launched." To see what it used to be like take a look at [the cached version](#).*

Launched in 2016 by “deep learning” group Youth Laboratories, [Beauty.AI](#) used age and facial recognition algorithms to choose what its creators declared would be “the First Beauty Queen or King Judged by Robots.” But the 44 winners – selected from among more than 7,000 entrants who had submitted selfies through the app – included only one with dark skin, although numerous people of colour had sent in photographs. The resulting media coverage saw the contest labelled as racist.

## Most Popular



Māori are trying to save their language from Big Tech

BY DONAVYN COFFEY



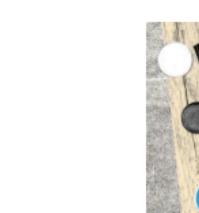
Why a James Bond film will never premiere on Netflix

BY WILL BEDINGFIELD



Here comes a new wave of green watches

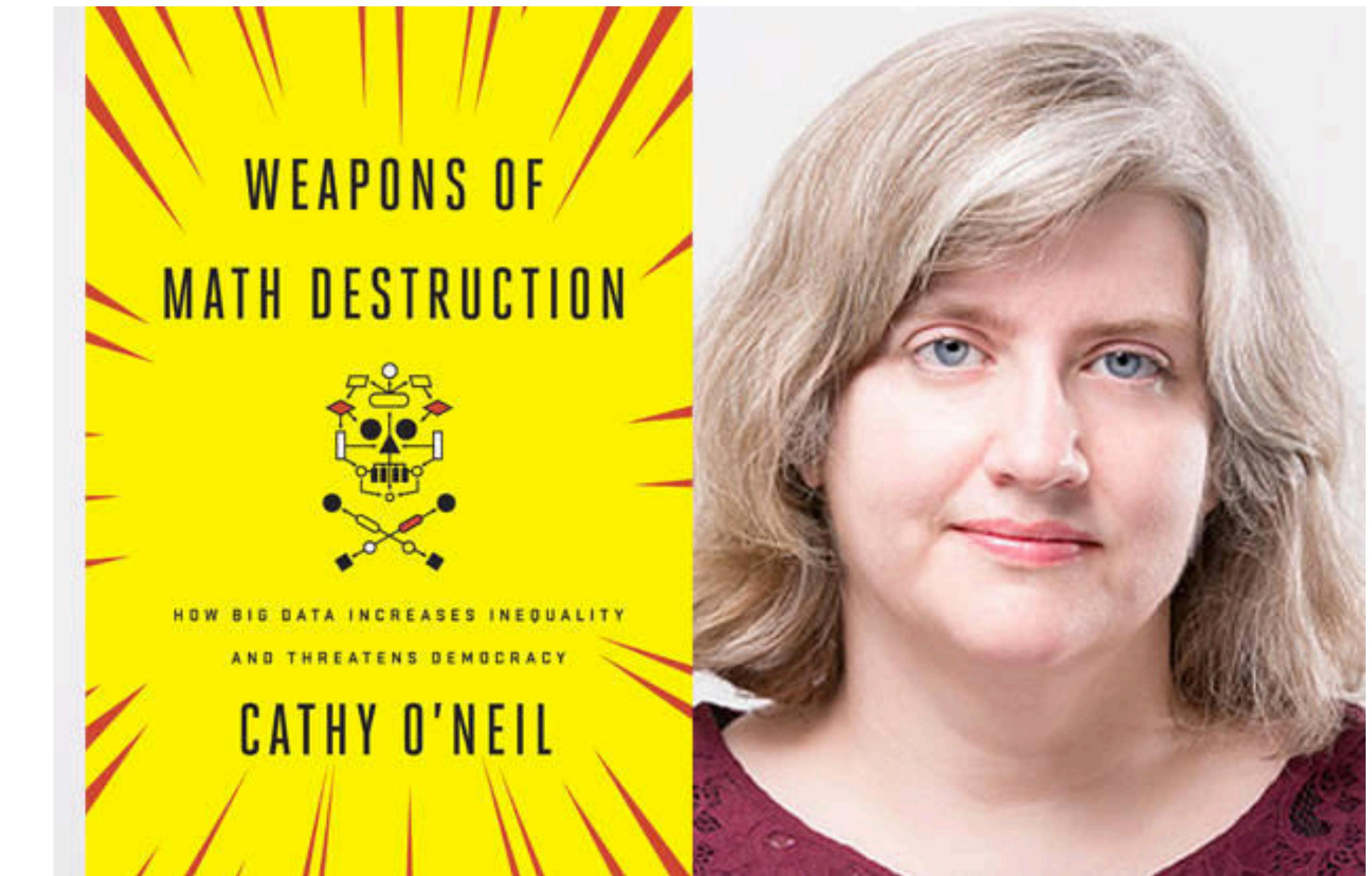
BY JEREMY WHITE



The future of social media is sharing less, not more

BY HUSSEIN KESVANI

## Cathy O'Neil: The era of blind faith in ~~big data~~<sup>black boxes</sup> must end



- “You don’t see a lot of skepticism,” she says. “The algorithms are like shiny new toys that we can’t resist using. We trust them so much that we project meaning on to them.”
- Ultimately algorithms, according to O’Neil, reinforce discrimination and widen inequality, “using people’s fear and trust of mathematics to prevent them from asking questions”.

<https://www.theguardian.com/books/2016/oct/27/cathy-oneil-weapons-of-math-destruction-algorithms-big-data>

What does it mean to discriminate?

## Key point

- Discrimination defines a situation where an individual is disadvantaged in some way on the basis of 'one or multiple protected grounds'.

HANDBOOK

## Handbook on European non-discrimination law

2018 edition



[https://fra.europa.eu/sites/default/files/fra\\_uploads/fra-2018-handbook-non-discrimination-law-2018\\_en.pdf](https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-handbook-non-discrimination-law-2018_en.pdf)

<https://fra.europa.eu/en/publication/2018/handbook-european-non-discrimination-law-2018-edition>

## Key point

- Discrimination defines a situation where an individual is disadvantaged in some way on the basis of 'one or multiple protected grounds'.

### PROTECTED GROUNDS

- Sex
- Gender identity
- Sexual orientation
- Disability
- Age
- Race, ethnicity, colour and membership of a national minority
- Nationality or national origin
- Religion or belief
- Social origin, birth and property
- Language
- Political or other opinion

[https://fra.europa.eu/sites/default/files/fra\\_uploads/fra-2018-handbook-non-discrimination-law-2018\\_en.pdf](https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-handbook-non-discrimination-law-2018_en.pdf)

HANDBOOK

**Handbook on European  
non-discrimination law**

2018 edition



<https://fra.europa.eu/en/publication/2018/handbook-european-non-discrimination-law-2018-edition>



Moritz Hardt

# Legally recognized ‘protected classes’ in the US

**Race** (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

# Principles for AI Ethics



## 1. Fairness

The company will strive to apply the values of **equality and diversity** in AI system throughout its entire lifecycle.

The company will strive not to **reinforce nor propagate negative or unfair bias**.

The company will strive to provide **easy access** to all users.



## 2. Transparency

Users will be aware that they are **interacting with AI**.

**AI will be explainable** for users to understand its decision or recommendation to the extent technologically feasible.

The process of collecting or utilizing personal data will be **transparent**.



## 3. Accountability

The company will strive to apply the principles of **social and ethical responsibility**.

AI system will be **adequately protected** and have **security measures** to prevent data breaches.

The company will strive to **benefit the society and promote the corporate citizen**.

## Trusted AI Lifecycle through Open Source

Pillars of trust, woven into the lifecycle of an AI application



Did anyone tamper with it?



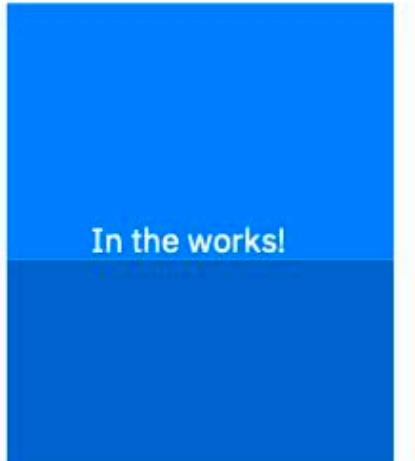
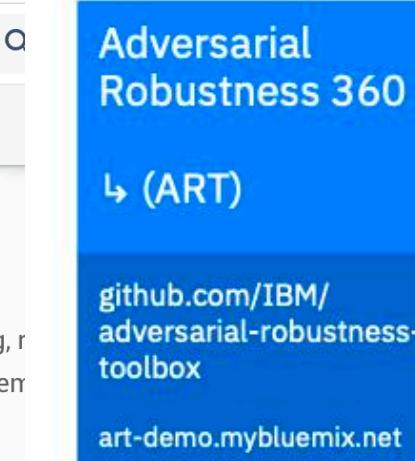
Is it fair?



Is it easy to understand?



Is it accountable?



### Recommended best practices for AI

Designing AI systems should follow software development best practices while taking a human-centered approach to ML

### Fairness

As the impact of AI increases across sectors and societies, it is critical to work towards systems that are fair and inclusive to everyone

### Interpretability

Understanding and trusting AI systems is important to ensuring they are working as intended

### Privacy

Training models off of sensitive data needs privacy preserving safeguards

### Security

Identifying potential threats can help keep AI systems safe and secure

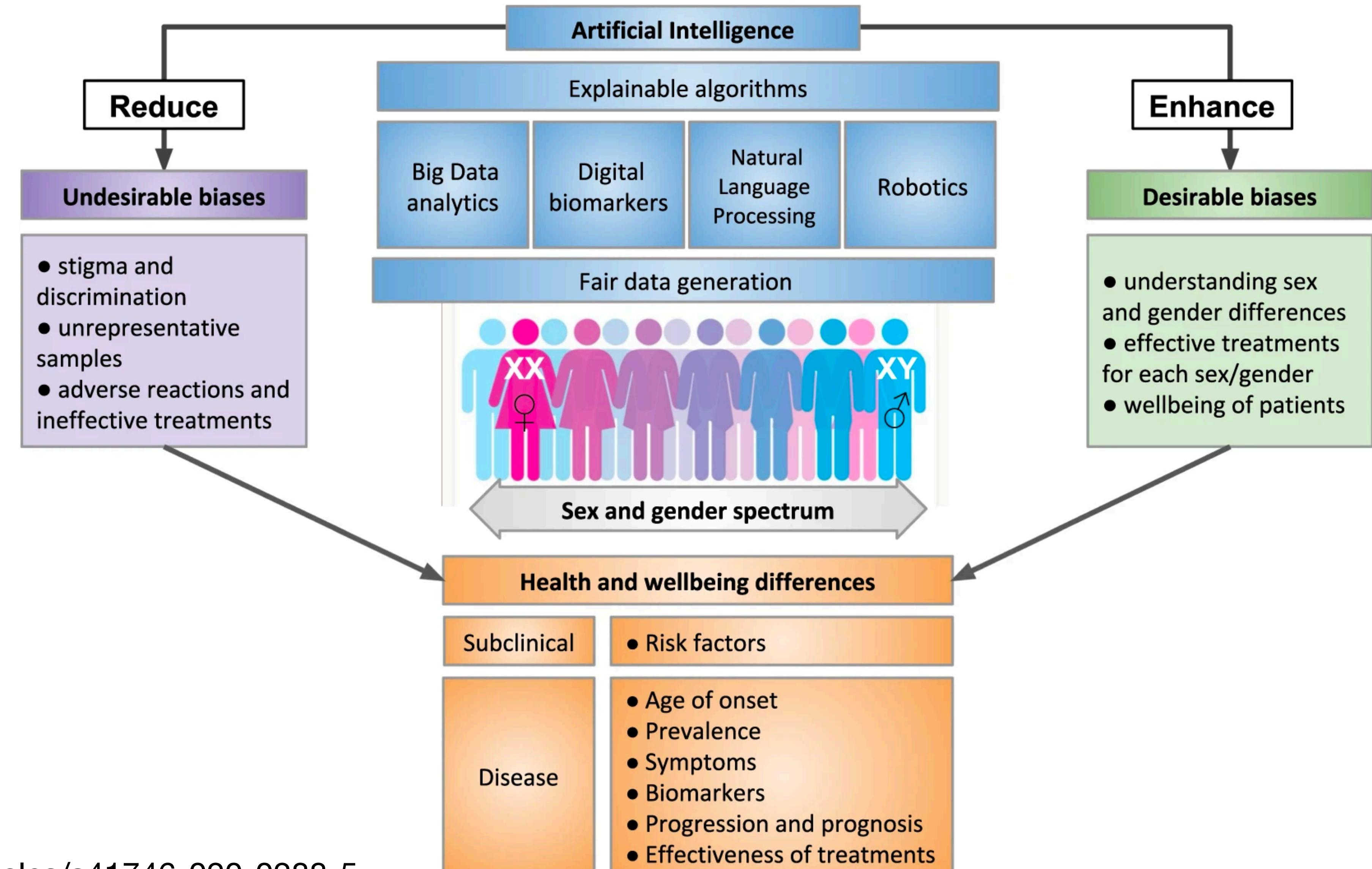
The screenshot shows the H2O.ai website with a yellow header bar. The main content area has a white background with a yellow title bar at the top. The title bar contains the H2O.ai logo, a search icon, and navigation links for Blog, Communities, Webinars & Events, Resources, Documentation, Downloads, and Contact Us. Below the title bar, there are several sections with icons and text. One section is titled "What is Responsible AI?" with a yellow background and a central yellow circle labeled "RESPONSIBLE AI". Surrounding this central circle are six smaller grey circles, each containing a different concept: Human-Centered Machine Learning, Explainable AI, Compliance, Ethical AI, Secure AI, and Interpretable Machine Learning. The overall design is clean and modern, using a color palette of yellow, grey, and white.

The screenshot shows the TensorFlow website with a white header bar. The main content area has a light grey background with a dark grey sidebar on the left. The sidebar contains the TensorFlow logo, a search icon, and navigation links for Install, Learn, API, Resources, Community, and Why TensorFlow. The main content area features a large yellow title bar with the text "What is Responsible AI?". Below the title bar, there is a grid of four blue boxes, each representing a different aspect of responsible AI: Adversarial Robustness 360, AI Fairness 360, AI Explainability 360, and Lineage. Each box contains a brief description, a link, and a small icon. To the right of the grid, there is a large black banner with white text that reads "A practical guide to Responsible Artificial Intelligence (AI)". The overall design is clean and professional, using a color palette of blue, white, and black.

Is different treatment always  
a discrimination?

# Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare

Cirillo et al 2020



# Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare

Cirillo et al 2020

Clinical conditions and studies	Current status without the desirable bias	Utility of the desirable bias
Autistic spectrum disorder	<p>There is a current lack of consideration of the demonstrated age-dependent sex differences in the symptomatology related with impairments in social communication and interaction, expressive behaviour, reciprocal conversation, non-verbal gestures for diagnostic purposes<sup>123</sup>.</p>	Differential diagnostic criteria for males and females could facilitate the identification of the clinical diagnosis leading to appropriate treatment.
Cardiovascular disorders	<p>Although it has been documented that men and women respond differently to many cardiovascular medications such as statins, angiotensin-converting enzyme inhibitors and β-Blockers among others, adopted treatments do not consider sex differences<sup>124</sup>.</p> <p>Despite the fact that Coronary heart disease (CHD) is the leading cause of death among women<sup>125</sup>, the majority (67%) of patients enroled in clinical trials for cardiovascular devices are male<sup>126</sup>.</p>	Making prescriptions according to the sex of the patient could lead to improved health benefits.
Genome-wide association studies (GWAS)	<p>Most of genome-wide association studies (GWAS) focus on white male subjects<sup>127</sup> and those that explore sex differences in complex traits are scarce<sup>128</sup>.</p>	The application of a desirable bias towards women would lead to a more accurate representation of sex differences in clinical research.
Human immunodeficiency virus (HIV)	<p>The observed lower female representation in HIV clinical trials depends, among other factors, from the disadvantaged awareness about treatment and enrolment options compared with men<sup>129–131</sup>.</p>	Promoting empowerment initiatives in those patients with disadvantages will increase their exposure to treatment options and clinical trial enrolment.

Think about the whole process  
bias may be everywhere

# Some sources of bias

- **Historical bias.** The data are correctly sampled and correspond well to the observed relationships, but due to different treatment in the past some prejudices are encoded in the data. Think about gender and occupation stereotypes.

# Some sources of bias

- **Historical bias.** The data are correctly sampled and correspond well to the observed relationships, but due to different treatment in the past some prejudices are encoded in the data. Think about gender and occupation stereotypes.
- **Representation bias.** The available data is not a representative sample of the population of interest. Think about the available facial images of actors, often white men. Or genetic sequences of covid variants, mostly collected in developed European countries. Or crime statistics in the regions to which the police are directed.

# Some sources of bias

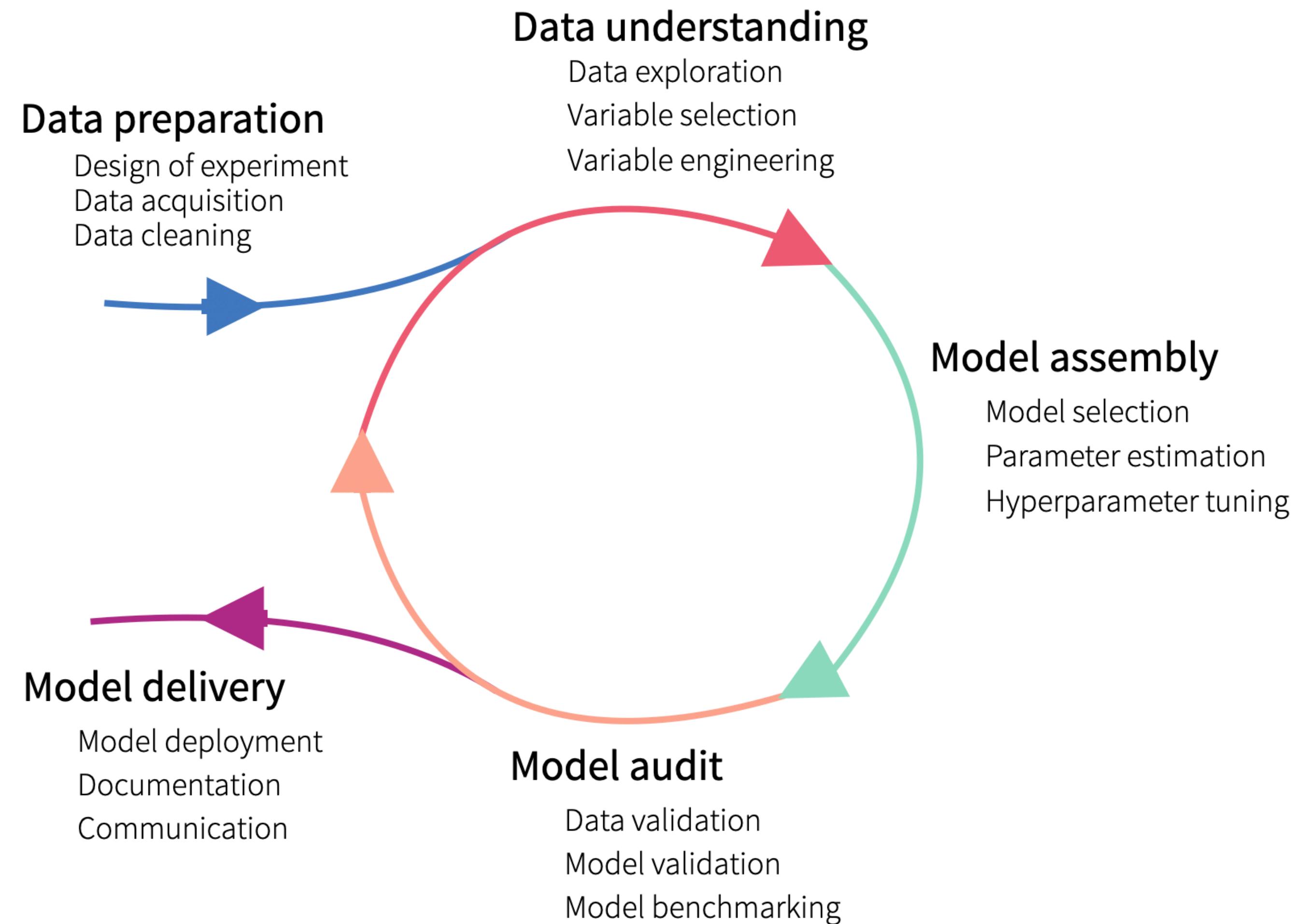
- **Historical bias.** The data are correctly sampled and correspond well to the observed relationships, but due to different treatment in the past some prejudices are encoded in the data. Think about gender and occupation stereotypes.
- **Representation bias.** The available data is not a representative sample of the population of interest. Think about the available facial images of actors, often white men. Or genetic sequences of covid variants, mostly collected in developed European countries. Or crime statistics in the regions to which the police are directed.
- **Measurement bias.** The variable of interest is not directly observable or is difficult to measure and the way it is measured may be distorted by other factors. Think of the results of the mathematics skills assessment (e.g. PISA) measured by tasks on computers not that widely available in some countries.

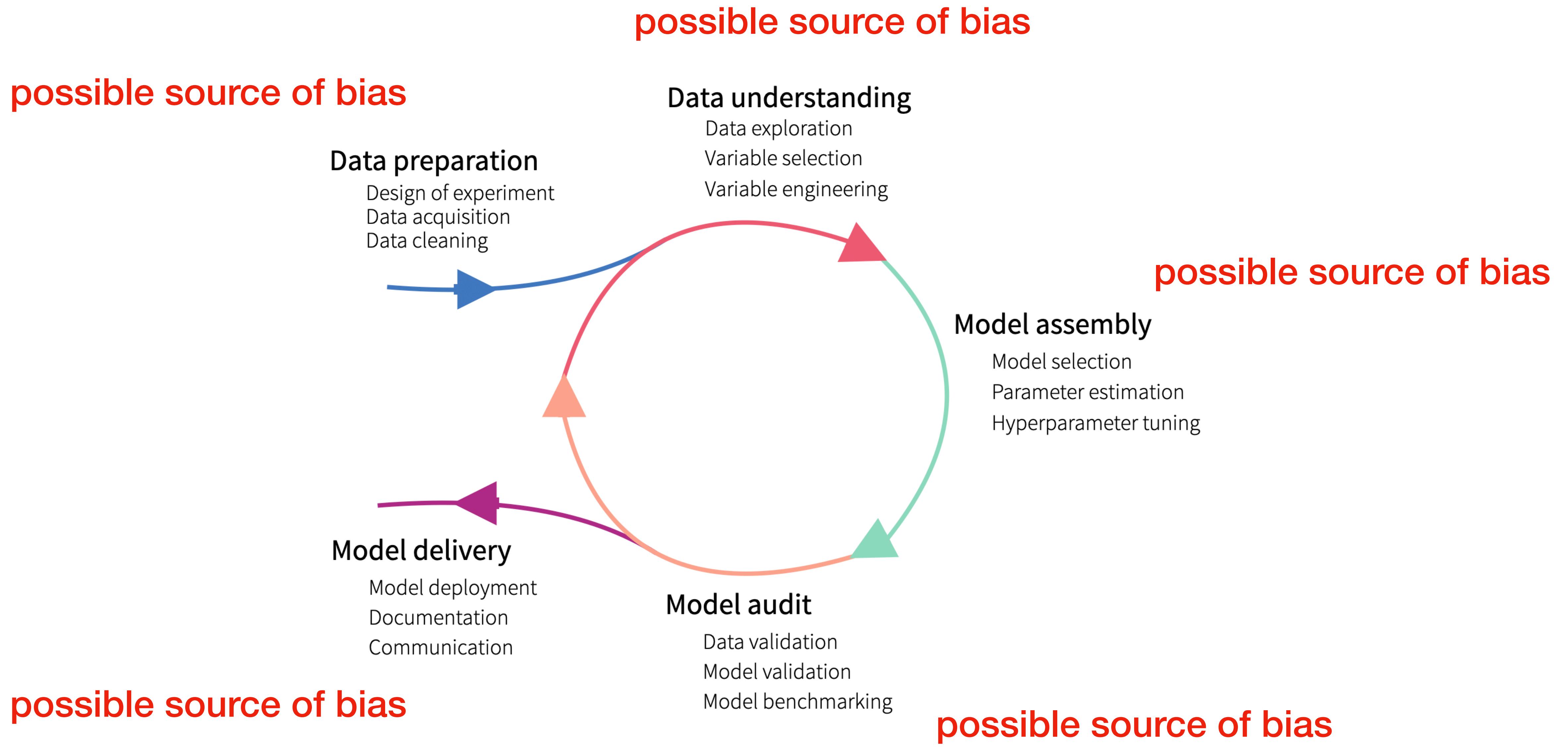
# Some sources of bias

- **Historical bias.** The data are correctly sampled and correspond well to the observed relationships, but due to different treatment in the past some prejudices are encoded in the data. Think about gender and occupation stereotypes.
- **Representation bias.** The available data is not a representative sample of the population of interest. Think about the available facial images of actors, often white men. Or genetic sequences of covid variants, mostly collected in developed European countries. Or crime statistics in the regions to which the police are directed.
- **Measurement bias.** The variable of interest is not directly observable or is difficult to measure and the way it is measured may be distorted by other factors. Think of the results of the mathematics skills assessment (e.g. PISA) measured by tasks on computers not that widely available in some countries.
- **Evaluation bias.** The evaluation of the algorithm is performed on a population that does not represent all groups. Think of a lung screening algorithm tested primarily on a population of smokers (older men).

# Some sources of bias

- **Historical bias.** The data are correctly sampled and correspond well to the observed relationships, but due to different treatment in the past some prejudices are encoded in the data. Think about gender and occupation stereotypes.
- **Representation bias.** The available data is not a representative sample of the population of interest. Think about the available facial images of actors, often white men. Or genetic sequences of covid variants, mostly collected in developed European countries. Or crime statistics in the regions to which the police are directed.
- **Measurement bias.** The variable of interest is not directly observable or is difficult to measure and the way it is measured may be distorted by other factors. Think of the results of the mathematics skills assessment (e.g. PISA) measured by tasks on computers not that widely available in some countries.
- **Evaluation bias.** The evaluation of the algorithm is performed on a population that does not represent all groups. Think of a lung screening algorithm tested primarily on a population of smokers (older men).
- **Proxy bias.** The algorithm uses variables that are proxies for protected attributes. Think of male/female only schools where the gender effect can be hidden under the school effect.





An interesting example is the StreetBumps project. The city of Boston released an application for mobile phones that allows to identify potholes based on vibrations measured by accelerometer. It is a very innovative idea, but when analyzing such data one has to take into account the representativeness of the collected data. Much more information about potholes will come from the neighborhoods where wealthier and younger people live, who use mobile phone more often.

# The Hidden Biases in Big Data

by Kate Crawford

April 01, 2013

This looks to be the year that we reach peak big data hype. From wildly popular big data conferences to columns in major newspapers, the business and science worlds are focused on how large datasets can give insight on previously intractable challenges. The hype becomes problematic when it leads to what I call “data fundamentalism,” the notion that correlation always indicates causation, and that massive data sets and predictive analytics always reflect objective truth. Former *Wired* editor-in-chief Chris Anderson embraced this idea in his comment, “with enough data, the numbers speak for themselves.” But can big data really deliver on that promise? Can numbers actually speak for themselves?

Sadly, they can’t. Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks, and are as important to the big-data equation as the numbers themselves.

Bias encoded in  
data embeddings

# Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Bolukbasi et al, 2016

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

- Extreme *she* occupations**
1. homemaker
  2. nurse
  3. receptionist
  4. librarian
  5. socialite
  6. hairdresser
  7. nanny
  8. bookkeeper
  9. stylist
  10. housekeeper
  11. interior designer
  12. guidance counselor

- Extreme *he* occupations**
1. maestro
  2. skipper
  3. protege
  4. philosopher
  5. captain
  6. architect
  7. financier
  8. warrior
  9. broadcaster
  10. magician
  11. fighter pilot
  12. boss

Figure 1: The most extreme occupations as projected on to the *she-he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

Gender stereotype <i>she-he</i> analogies.		
sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber
Gender appropriate <i>she-he</i> analogies.		
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 2: **Analogy examples.** Examples of automatically generated analogies for the pair *she-he* using the procedure described in text. For example, the first analogy is interpreted as *she:sewing :: he:carpentry* in the original w2vNEWS embedding. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype. Top: illustrative gender stereotypic analogies automatically generated from w2vNEWS, as rated by at least 5 of the 10 crowd-workers. Bottom: illustrative generated gender-appropriate analogies.

Measuring Bias in Contextualized Word Representations  
Bolukbasi et al 2016, <https://arxiv.org/pdf/1607.06520.pdf>

Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems  
Kiritchenko et al 2018, <https://arxiv.org/pdf/1805.04508.pdf>

Learning Gender-Neutral Word Embeddings  
Zhao et al 2018, <https://aclanthology.org/D18-1521/>

Mitigating Gender Bias in Natural Language Processing: Literature Review  
Sun et al 2019, <https://arxiv.org/pdf/1906.08976.pdf>

Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function  
Qian et al 2019, <https://arxiv.org/pdf/1905.12801.pdf>

Identifying and Reducing Gender Bias in Word-Level Language Models  
Bordia et al 2019, <https://aclanthology.org/N19-3002.pdf>

Investigating Gender Bias in Language Models Using Causal Mediation Analysis  
Vig et al 2020, <https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf>

Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models  
de Vassimon Manela et al 2021, <https://arxiv.org/pdf/2101.09688.pdf>

Is it always about higher  
scores?

BUSINESS NEWS

OCTOBER 10, 2018 / 5:12 AM / A MONTH AGO

## Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN

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

The group created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes. The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes, the people said.

Instead, the technology favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured," one person said.

**Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist**

THE VERGE

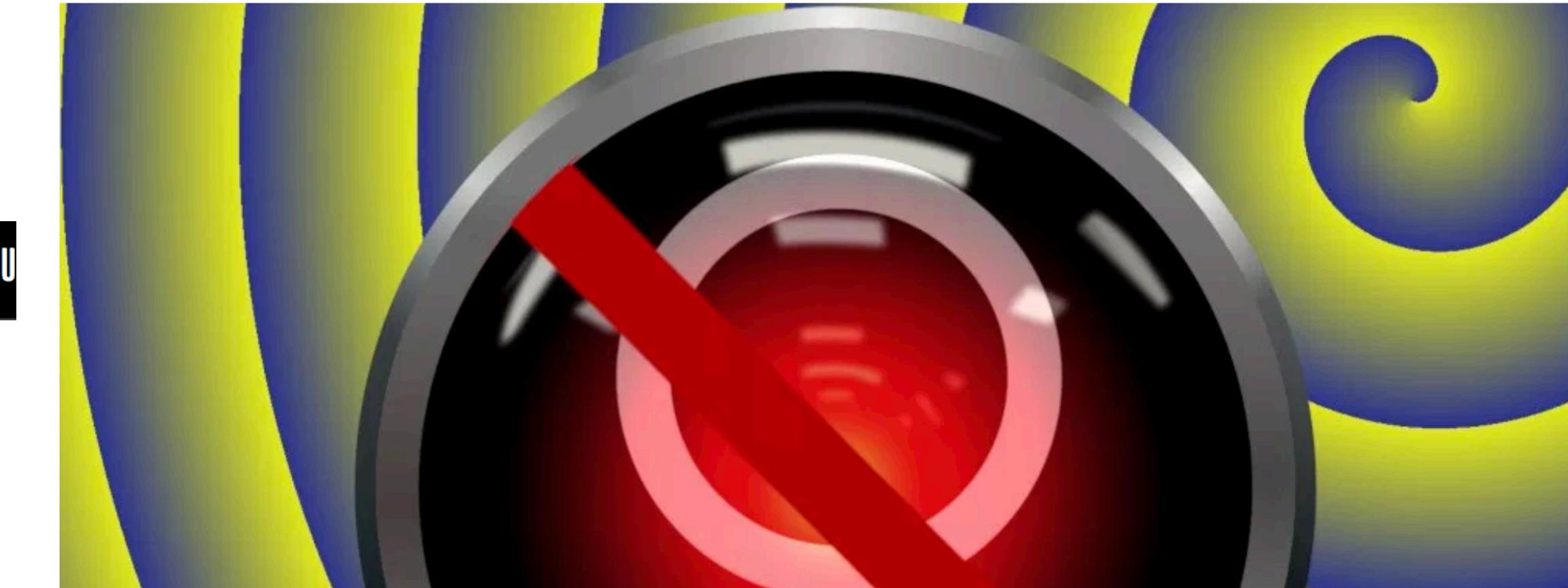
TECH ▾ SCIENCE ▾ CULTURE

TECH ▾ AMAZON ▾ ARTIFICIAL INTELLIGENCE

## Amazon reportedly scraps internal AI recruiting tool that was biased against women

21

*The secret program penalized applications that contained the word "women's"*

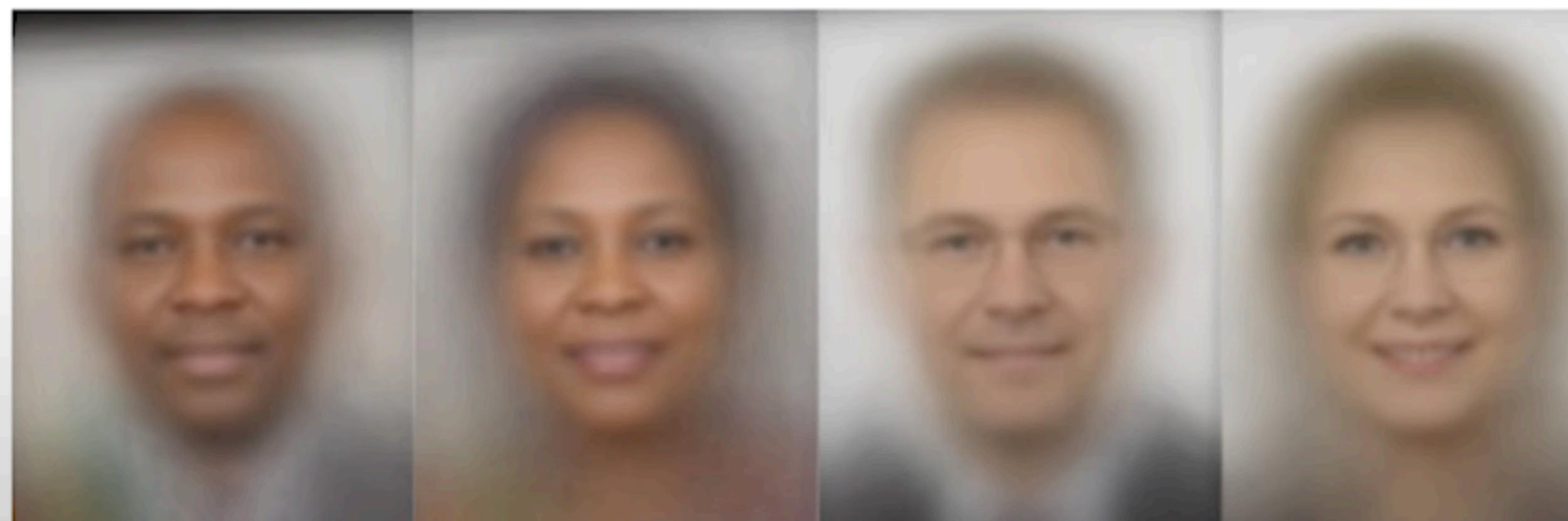


# Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

Buolamwini and Gebru, 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%

Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	100
Error Rate(%)	6.3	10.7	2.6	12.9	0.7	<b>20.8</b>	6.0	1.7	0.0
TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	<b>100</b>	98.7
FPR (%)	6.3	8.3	3.5	12.9	0.7	<b>16.3</b>	7.9	1.3	0.0
PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	<b>99.3</b>	94.0	99.2
Error Rate(%)	10.0	21.3	0.7	16.5	4.7	<b>34.5</b>	0.7	6.0	0.8
TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	<b>98.9</b>	92.9
FPR (%)	10.0	14.9	1.1	16.5	4.7	<b>23.4</b>	1.2	7.1	1.1
PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	<b>99.7</b>
Error Rate(%)	12.1	20.3	5.6	22.4	3.2	<b>34.7</b>	12.0	7.1	0.3
TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	<b>99.6</b>	94.8
FPR (%)	12.1	14.8	7.9	22.4	3.2	<b>25.2</b>	17.7	5.20	0.4



# Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

Buolamwini and Gebru, 2018

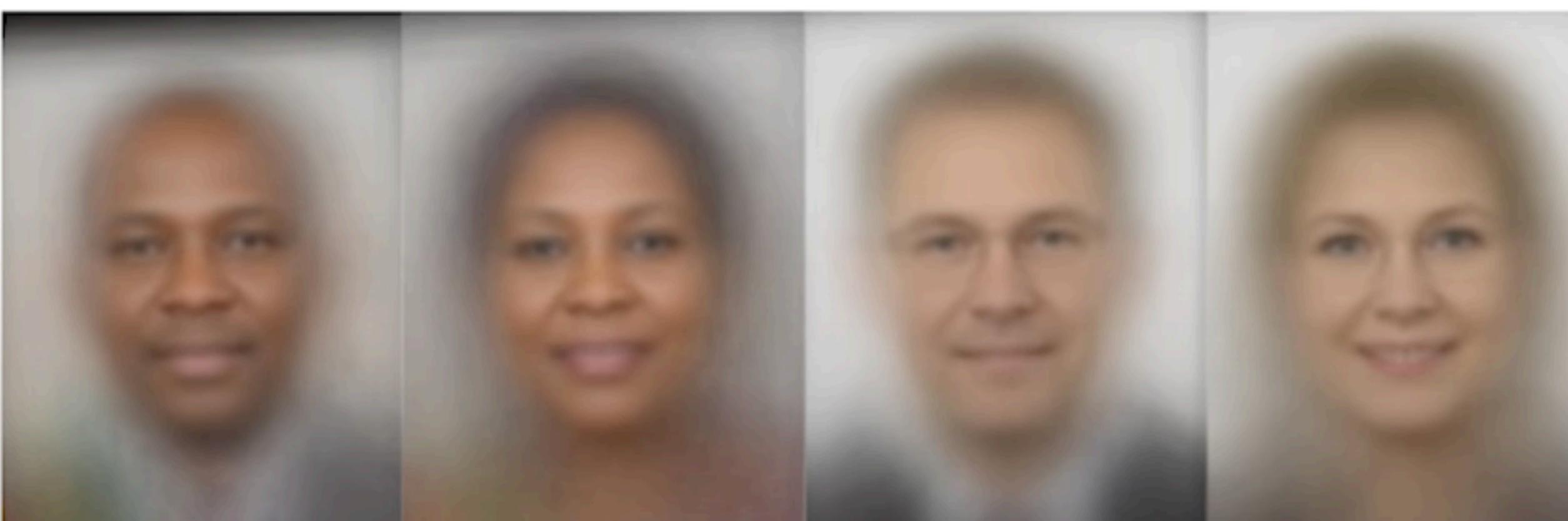
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Lower performance in a group can generate a problem in future data.

Think about credit lines.

If in group A loans are assigned with high accuracy and in group B loans are assigned with low accuracy then over time we will get a dataset where in group A many loans are paid and in group B not paid.

This is only because the people who were assigned credits in group B were more random.



Disregarding sensitive  
attributes is not a solution

# Apple's credit card is being investigated for discriminating against women

Customers say the card offers less credit to women than men

By James Vincent | Nov 11, 2019, 5:57am EST

If you buy something from a Verge link, Vox Media may earn a commission. See our [ethics statement](#).

[f](#) [t](#) [s](#) SHARE



Image: Apple

Apple's new [credit card](#) is being investigated by financial regulators after customers complained that the card's lending algorithms discriminated against women.

117 [P](#)



**DHH** @dhh · Nov 9, 2019

Replying to @dhh

To be fair, this is an even more egregious version of the same take. THE ALGORITHM is always assumed to be just and correct. It's verdict is thus predestined to be a reflection of your failings and your sins.



**Steve Wozniak** @stevewoz · Nov 10, 2019

The same thing happened to us. We have no separate bank accounts or credit cards or assets of any kind. We both have the same high limits on our cards, including our AmEx Centurion card. But 10x on the Apple Card.

19

86

1.4K

↑



**Isles47** @isles47 · Nov 9, 2019

Replying to @dhh and @AppleCard

Haha this is absurd. Literally none of the things you list here have any effect on credit approval. What's her existing line of credit, what's her credit score, what outstanding debt does she have? How old is her original line of credit?

19

86

1.4K

↑

33

68

182

↑

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.



Image: Apple

Apple's new [credit card](#) is being investigated by financial regulators after customers complained that the card's lending algorithms discriminated against women.



**Steve Wozniak** @stevewoz · Nov 10, 2019

The same thing happened to us. We have no separate bank accounts or credit cards or assets of any kind. We both have the same high limits on our cards, including our AmEx Centurion card. But 10x on the Apple Card.

33

68

182

↑

The risk of  
„gaming the fairness”

## Computer Science &gt; Machine Learning

*[Submitted on 20 Jul 2020]*

# Fairwashing Explanations with Off-Manifold Detergent

**Christopher J. Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, Pan Kessel**

Explanation methods promise to make black-box classifiers more transparent. As a result, it is hoped that they can act as proof for a sensible, fair and trustworthy decision-making process of the algorithm and thereby increase its acceptance by the end-users. In this paper, we show both theoretically and experimentally that these hopes are presently unfounded. Specifically, we show that, for any classifier  $g$ , one can always construct another classifier  $\tilde{g}$  which has the same behavior on the data (same train, validation, and test error) but has arbitrarily manipulated explanation maps. We derive this statement theoretically using differential geometry and demonstrate it experimentally for various explanation methods, architectures, and datasets. Motivated by our theoretical insights, we then propose a modification of existing explanation methods which makes them significantly more robust.

Comments: 22 pages with 43 figures, to be published in ICML2020

Subjects: **Machine Learning (cs.LG); Machine Learning (stat.ML)**

Cite as: [arXiv:2007.09969 \[cs.LG\]](#)

(or [arXiv:2007.09969v1 \[cs.LG\]](#) for this version)

## Submission history

From: Christopher J. Anders [[view email](#)]

[v1] Mon, 20 Jul 2020 09:42:06 UTC (4,823 KB)

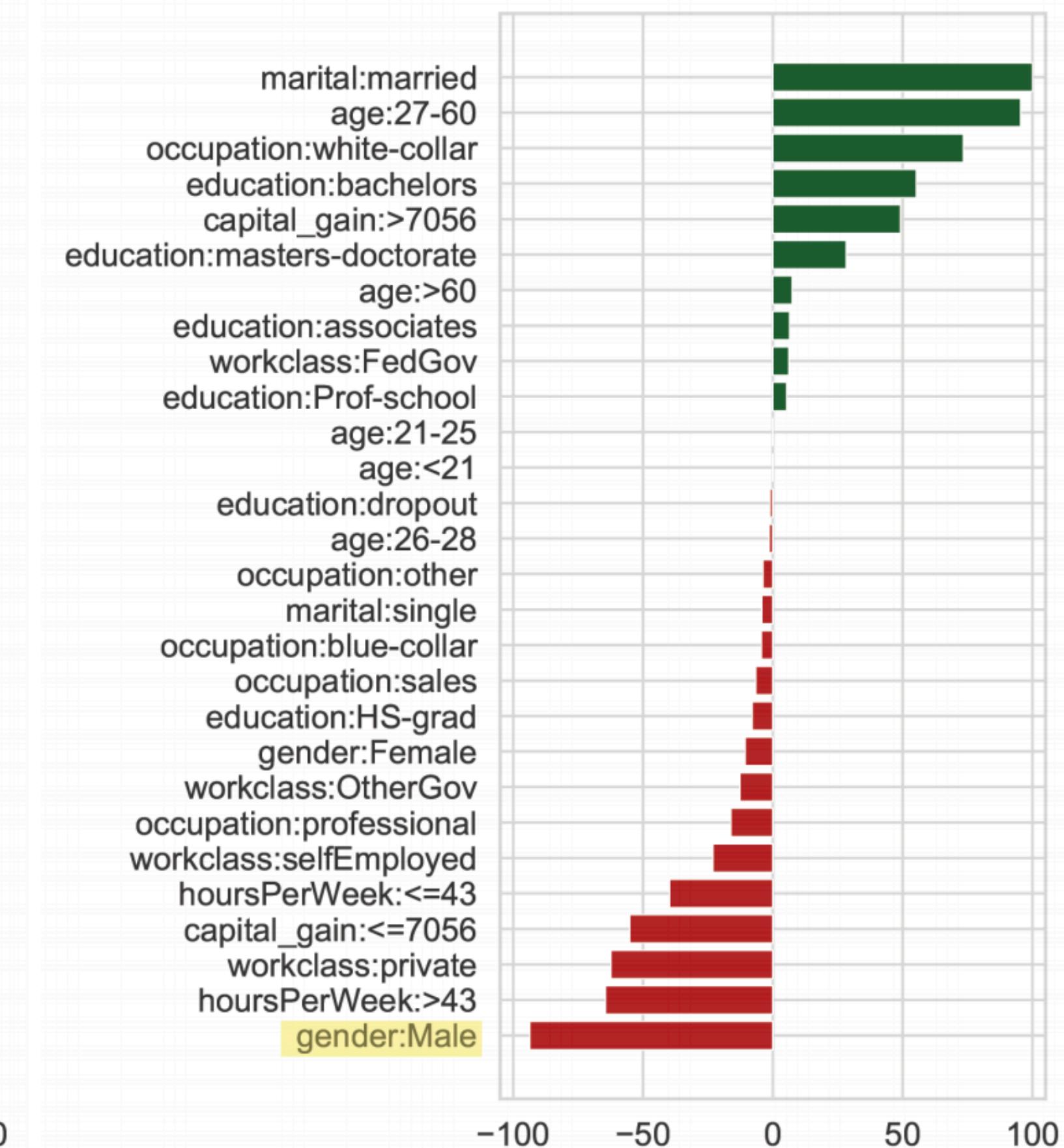
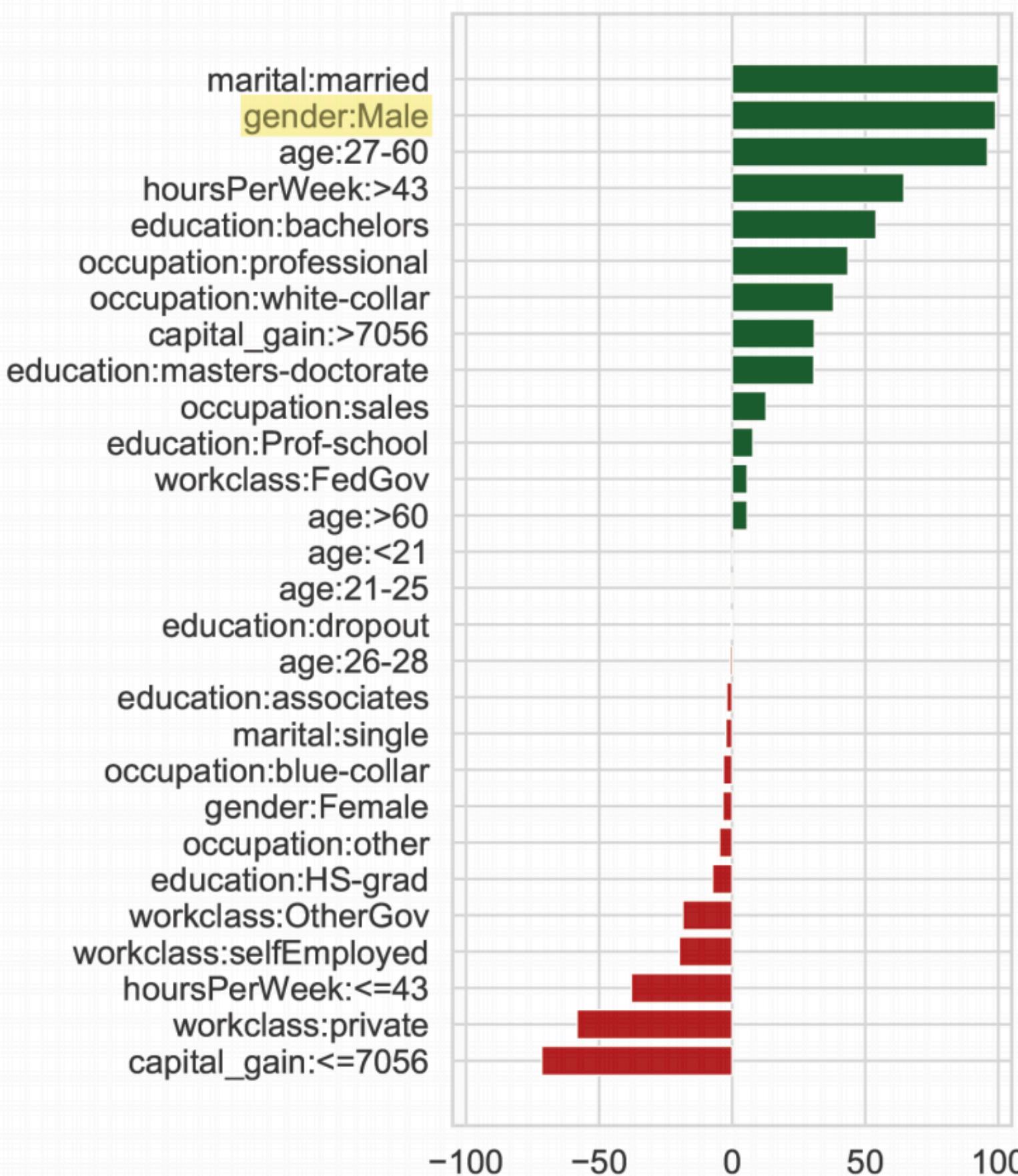
<https://arxiv.org/abs/2007.09969>

# Fairwashing: the risk of rationalization

Ulrich Aïvodji<sup>1</sup> Hiromi Arai<sup>2,3</sup> Olivier Fortineau<sup>4</sup> Sébastien Gambs<sup>1</sup> Satoshi Hara<sup>5</sup> Alain Tapp<sup>6,7</sup>

## Abstract

Black-box explanation is the problem of explaining how a machine learning model – whose internal logic is hidden to the auditor and generally complex – produces its outcomes. Current approaches for solving this problem include model explanation, outcome explanation as well as model inspection. While these techniques can be beneficial by providing interpretability, they can be used in a negative manner to perform fairwashing, which we define as promoting the false perception that a machine learning model respects some ethical values. In particular, we demonstrate that it is possible to systematically rationalize decisions taken by an unfair black-box model using the model explanation as well as the outcome explanation approaches with a given fairness metric. Our solution, LaundryML, is based on a regularized rule list enumeration algorithm whose objective is to search for fair rule lists approximating an unfair black-box model. We empirically evaluate our rationalization technique on black-box models trained on real-world datasets and show that one can obtain rule lists with high fidelity to the black-box model while being considerably less unfair at the same time.



# Fairness measures

# Notation

$Y$

true class, (1 - is preferred, favourable outcome)

$S$

predicted score

$\hat{Y} := 1_{S>c}$

decision

$A$

protected attribute

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

# Group fairness / statistical parity / independence / demographic parity

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = b) \quad \hat{Y} \perp A$$

Predicted class is independent from protected attribute

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

# Group fairness / statistical parity / independence / demographic parity

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = b) \quad \hat{Y} \perp A$$

Predicted class is independent from protected attribute

„four-fifth rule“ - selection rate for any disadvantaged group that is less than four-fifths of that for the group with the highest rate

$$80\% \leq \frac{P(\hat{Y}=1|A=a)}{P(\hat{Y}=1|A=b)} \leq 125\%$$

# Group fairness / statistical parity / independence / demographic parity

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = b) \quad \hat{Y} \perp A$$

Predicted class is independent from protected attribute

„four-fifth rule“ - selection rate for any disadvantaged group that is less than four-fifths of that for the group with the highest rate

$$80\% \leq \frac{P(\hat{Y}=1|A=a)}{P(\hat{Y}=1|A=b)} \leq 125\%$$

Sounds like a good idea, but is easy to fool.

For example, in class a we use a valid classifier while in class b we make decisions randomly.

Perfect classifier does not satisfy this parity if classes are not balanced.

# Equal opportunity

$$P(\hat{Y} = 1 | A = a, Y = 1) = P(\hat{Y} = 1 | A = b, Y = 1)$$

Equal True Positive Rate  $\text{TPR} = \text{TP}/(\text{TP}+\text{FN})$  for each subgroup

If he/she pays the credit then should have an equal chance to get it.

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

# Predictive equality

$$P(\hat{Y} = 1|A = a, Y = 0) = P(\hat{Y} = 1|A = b, Y = 0)$$

Equal False Positive Rate  $FPR = FP/(FP+TN)$  for each subgroup

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

# Equalized odds, Separation, Positive Rate Parity

$$P(\hat{Y} = 1|A = a, Y = 1) = P(\hat{Y} = 1|A = b, Y = 1)$$

$$\hat{Y} \perp A \mid Y$$

$$P(\hat{Y} = 1|A = a, Y = 0) = P(\hat{Y} = 1|A = b, Y = 0)$$

Equal True Positive Rate  $\text{TPR} = \text{TP}/(\text{TP}+\text{FN})$  for each subgroup and

equal False Positive Rate  $\text{FPR} = \text{FP}/(\text{FP}+\text{TN})$  for each subgroup

Predicted class is independent from protected attribute given true class

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

# Positive Predictive Parity

$$P(Y = 1 | A = a, \hat{Y} = 1) = P(Y = 1 | A = b, \hat{Y} = 1)$$

Equal Positive Predictive Value  $\text{PPV} = \text{TP} / (\text{TP} + \text{FP})$  for each subgroup

If he/she gets the credit then should have an equal chance to pay it.

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

# Negative Predictive Parity

$$P(Y = 1|A = a, \hat{Y} = 0) = P(Y = 1|A = b, \hat{Y} = 0)$$

Equal Negative Predictive Value  $NPV = NP / (NP + FN)$  for each subgroup

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

# Predictive Rate Parity, Sufficiency

$$P(Y = 1|A = a, \hat{Y} = 1) = P(Y = 1|A = b, \hat{Y} = 1) \quad Y \perp A | \hat{Y}$$

$$P(Y = 1|A = a, \hat{Y} = 0) = P(Y = 1|A = b, \hat{Y} = 0)$$

Equal Positive Predictive Value  $PPV = TP / (TP + FP)$  for each subgroup and

equal Negative Predictive Value  $NPV = NP / (NP + FN)$  for each subgroup

True class is independent from protected attribute given predicted class

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

# Whether to sentence a prisoner

**Demographic parity:**

the rate of sentenced prisoners should be equal in each group

*fair from society's perspective*

**Equal opportunity:**

The fraction of innocents sentenced should be equal in subgroups

*fair from the prisoner's perspective (ProPublica)*

**Predictive Rate Parity:**

Among the convicted, there should be an equal fraction of innocents

*fair from the judge's perspective (Northpointe)*

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

	$Y = 1$	$Y = 0$	
$\hat{Y} = 1$	TP	FP	P
$\hat{Y} = 0$	FN	TN	N
	$\pi_1$	$\pi_0$	

# The Fairness Trade-off (the impossibility theorem)

Except for trivial cases all these criteria cannot be satisfied jointly.

In fact each two out of {Sufficiency, Separation, Independence} are mutually exclusive.

# Bias mitigation strategies

## **Data Pre-processing**

change data to improve model performance, for example, use subsampling or case weighting

## **Model In-processing**

modify the optimized criterion to include fairness functions, e.g. through adversarial training

## **Model Post-processing**

modify the resulting model scores or final decisions, e.g., using different thresholds

Hands on example  
with the fairmodels package

**Statistics > Machine Learning***[Submitted on 1 Apr 2021]*

# fairmodels: A Flexible Tool For Bias Detection, Visualization, And Mitigation

**Jakub Wiśniewski, Przemysław Biecek**

Machine learning decision systems are getting omnipresent in our lives. From dating apps to rating loan seekers, algorithms affect both our well-being and future. Typically, however, these systems are not infallible. Moreover, complex predictive models are really eager to learn social biases present in historical data that can lead to increasing discrimination. If we want to create models responsibly then we need tools for in-depth validation of models also from the perspective of potential discrimination. This article introduces an R package fairmodels that helps to validate fairness and eliminate bias in classification models in an easy and flexible fashion. The fairmodels package offers a model-agnostic approach to bias detection, visualization and mitigation. The implemented set of functions and fairness metrics enables model fairness validation from different perspectives. The package includes a series of methods for bias mitigation that aim to diminish the discrimination in the model. The package is designed not only to examine a single model, but also to facilitate comparisons between multiple models.

Comments: 15 pages, 9 figures

Subjects: **Machine Learning (stat.ML)**; Machine Learning (cs.LG); Mathematical Software (cs.MS); Applications (stat.AP)

Cite as: [arXiv:2104.00507 \[stat.ML\]](#)

(or [arXiv:2104.00507v1 \[stat.ML\]](#) for this version)

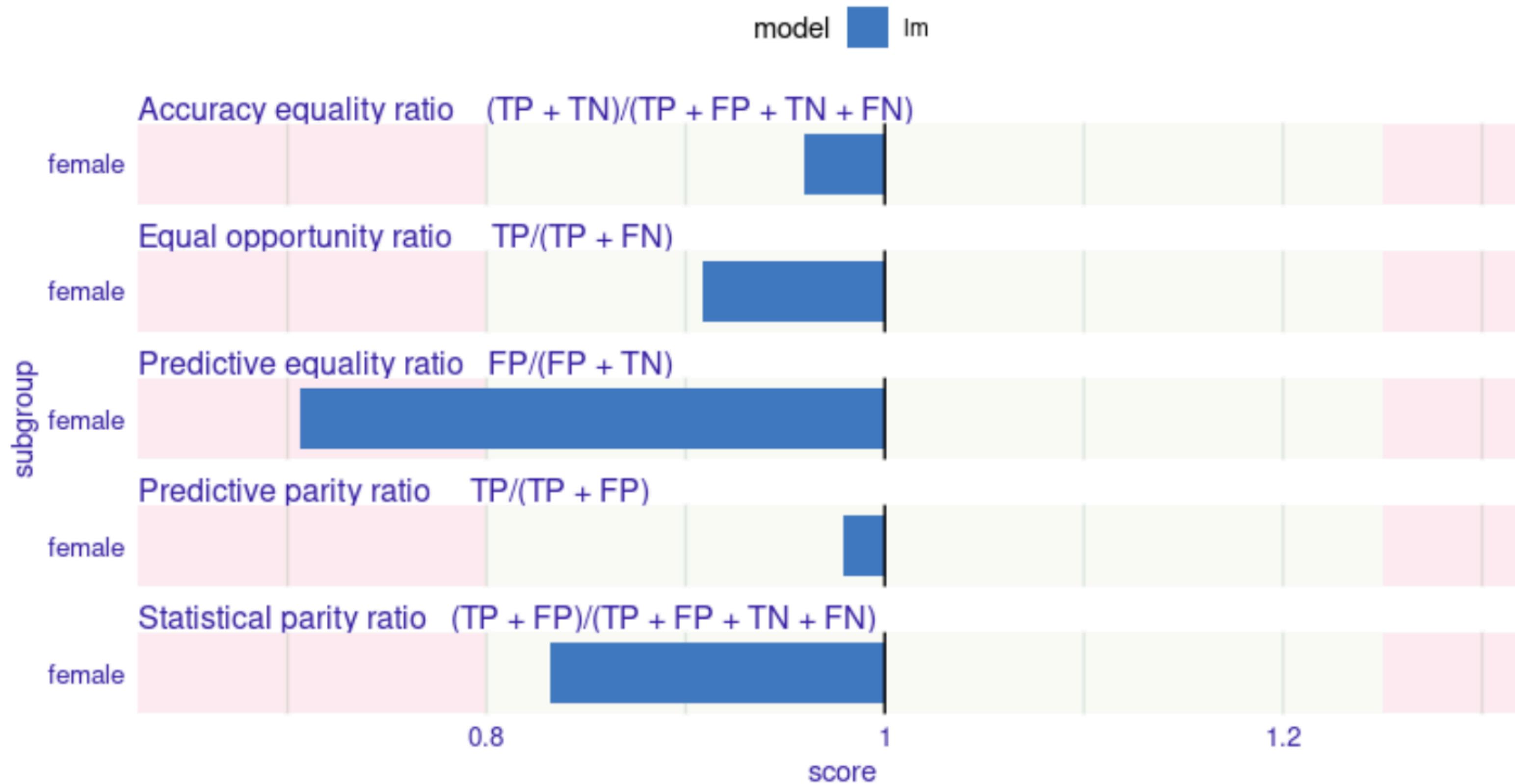
## Submission history

From: Przemysław Biecek [[view email](#)]

[v1] Thu, 1 Apr 2021 15:06:13 UTC (1,920 KB)

# Fairness check

Created with lm

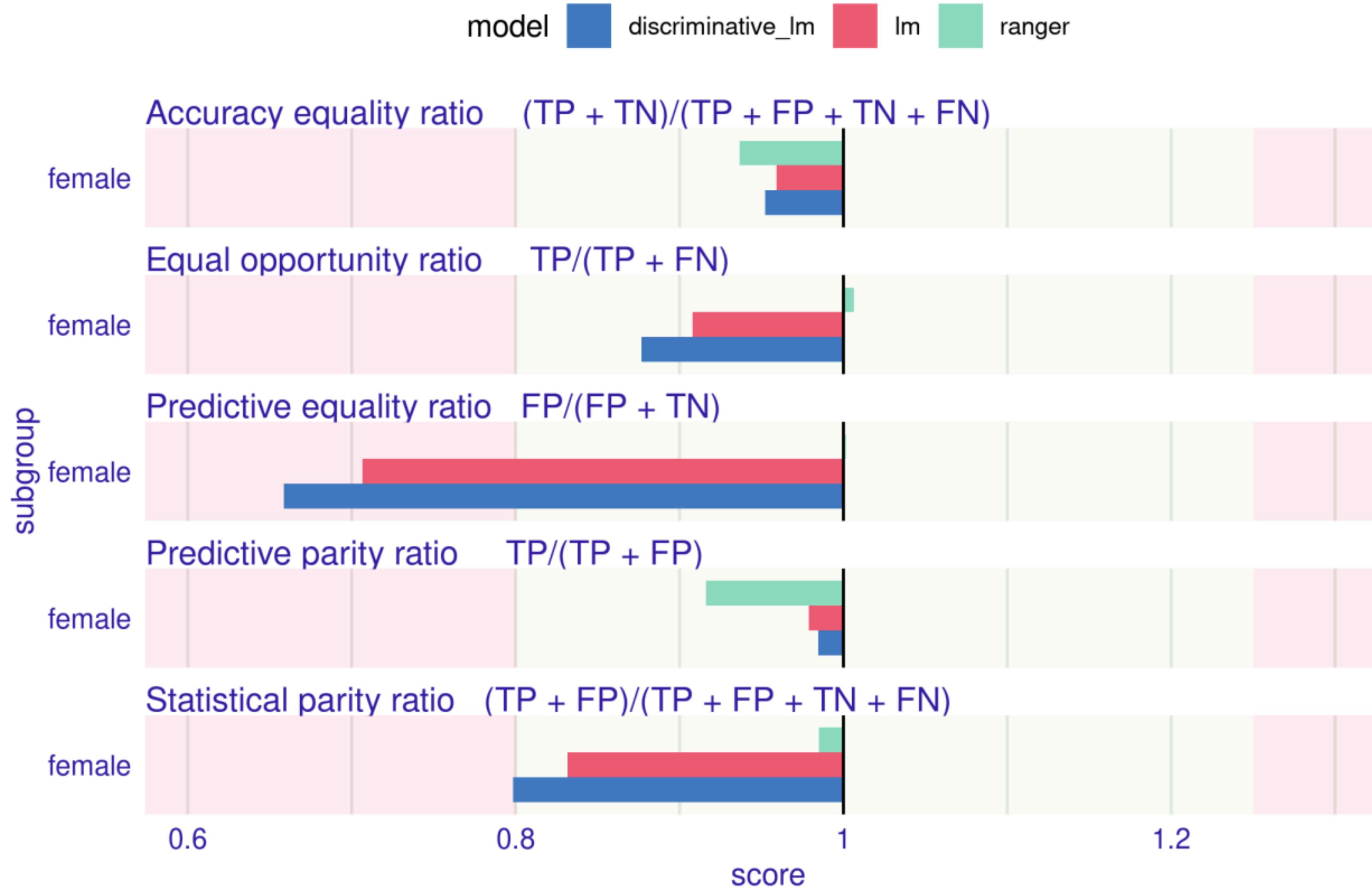


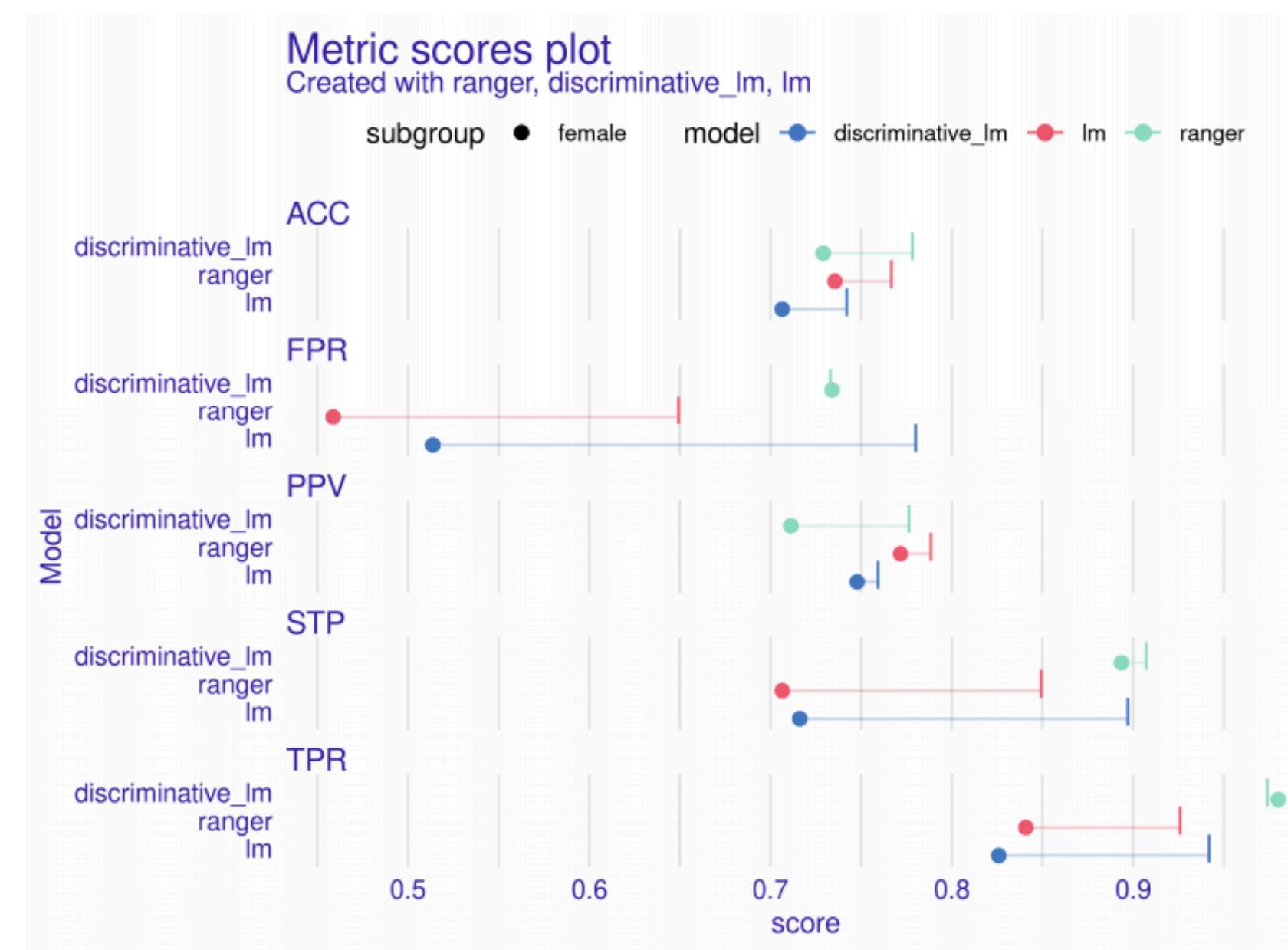
The four-fifths rule (Code of Federal Regulations, 1978)

"A selection rate for any race, sex, or ethnic group which is less than four-fifths (4 / 5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact[...]."

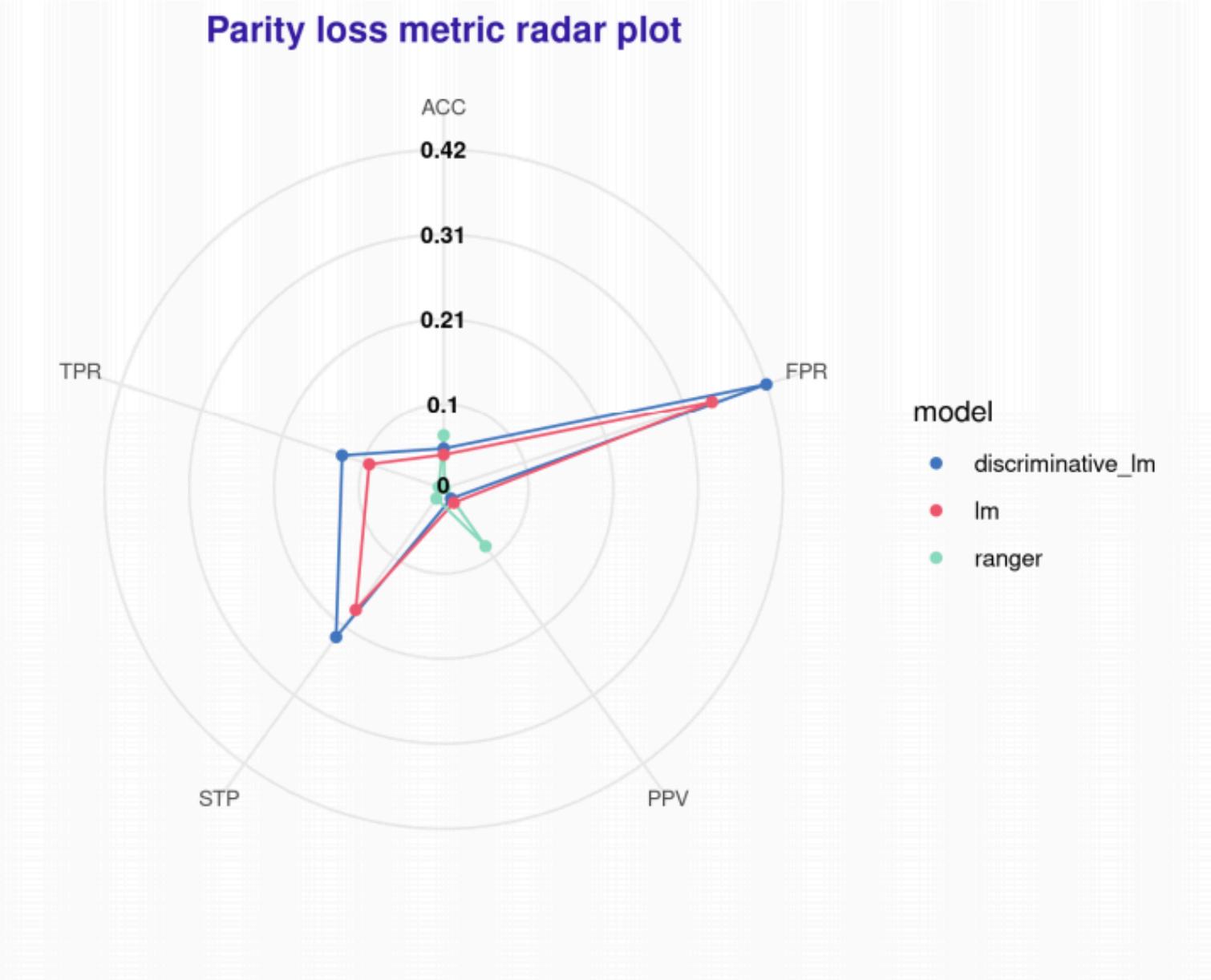
# Fairness check

Created with ranger, discriminative\_lm, lm

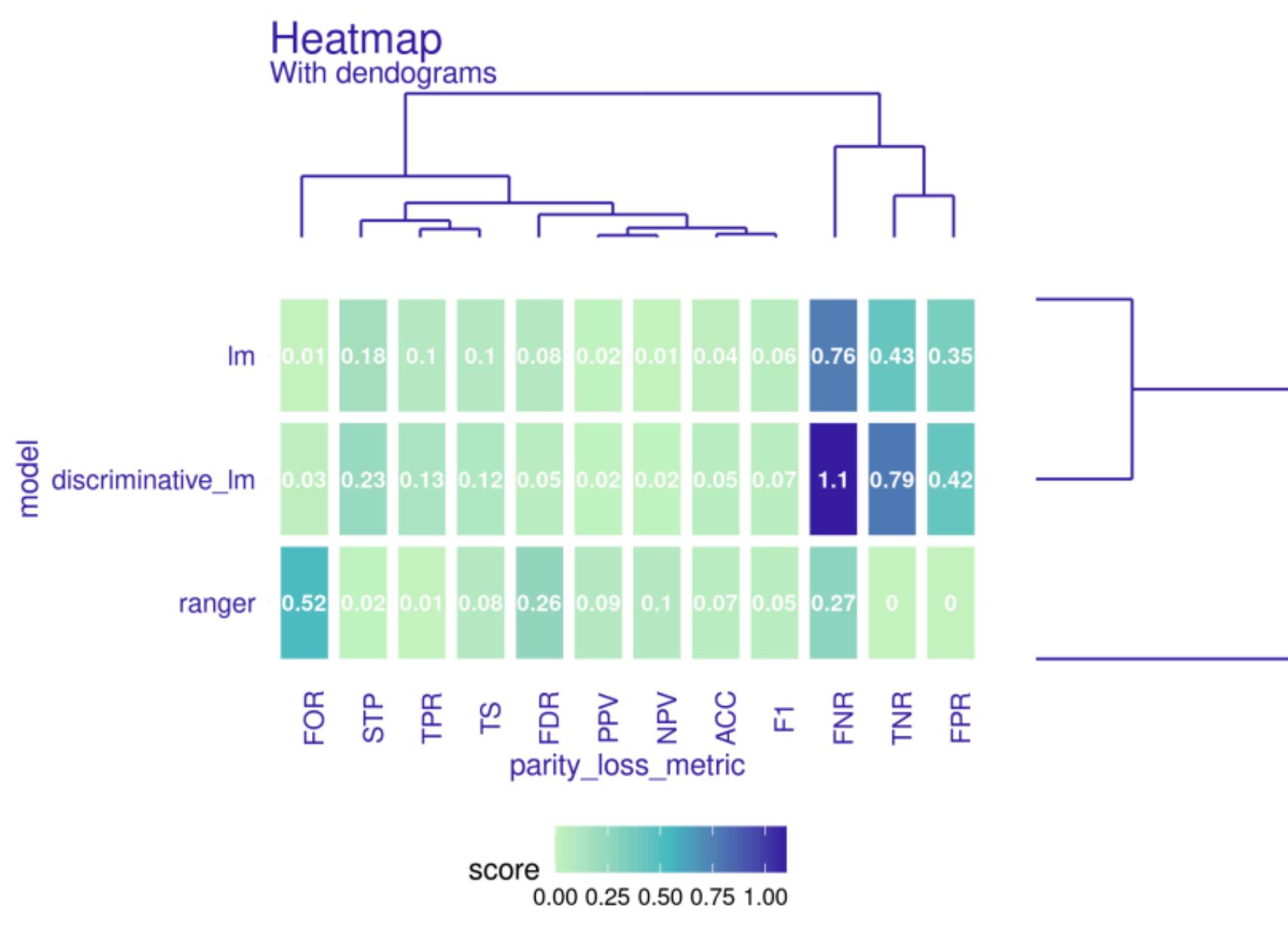




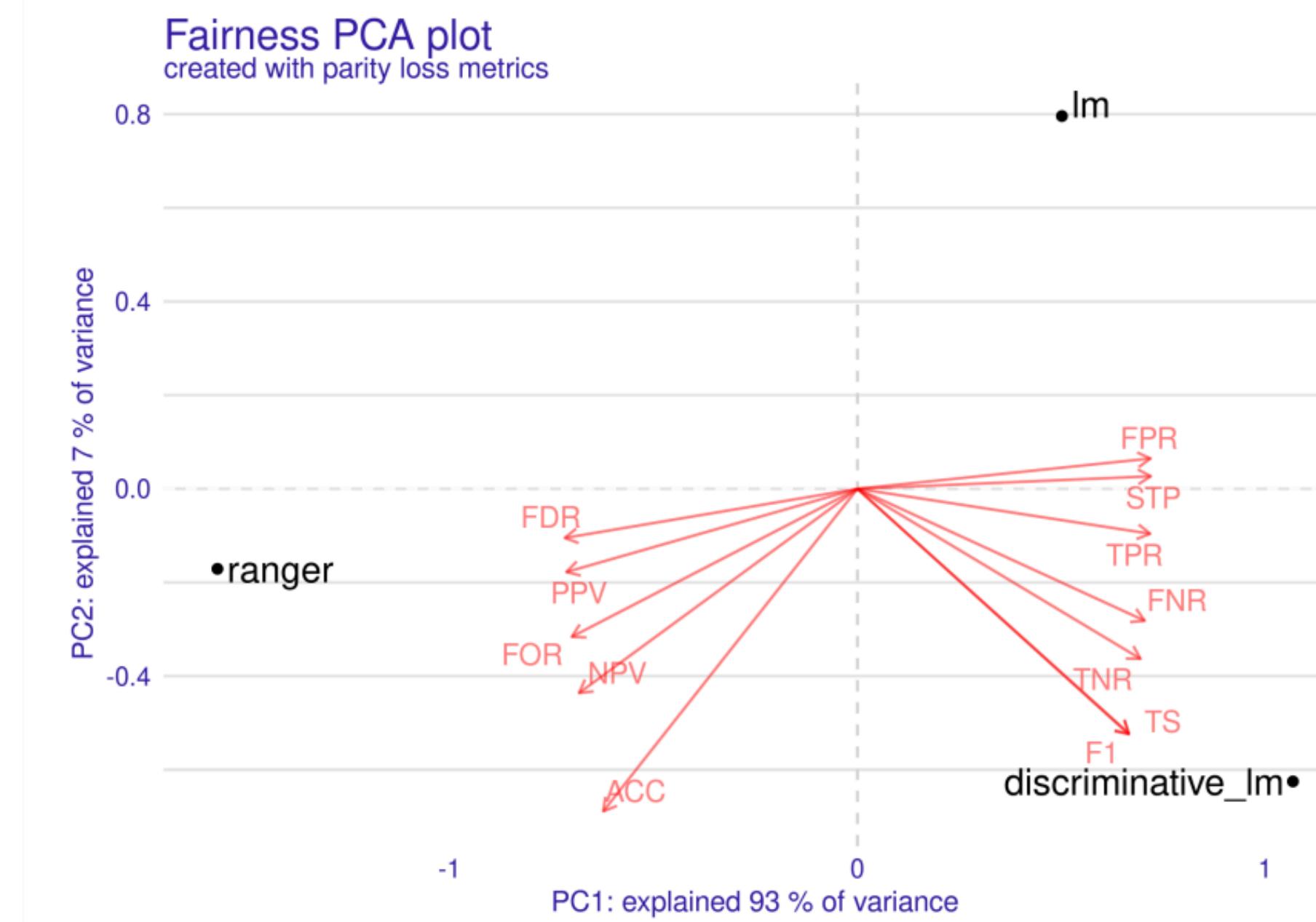
(a) metric\_scores



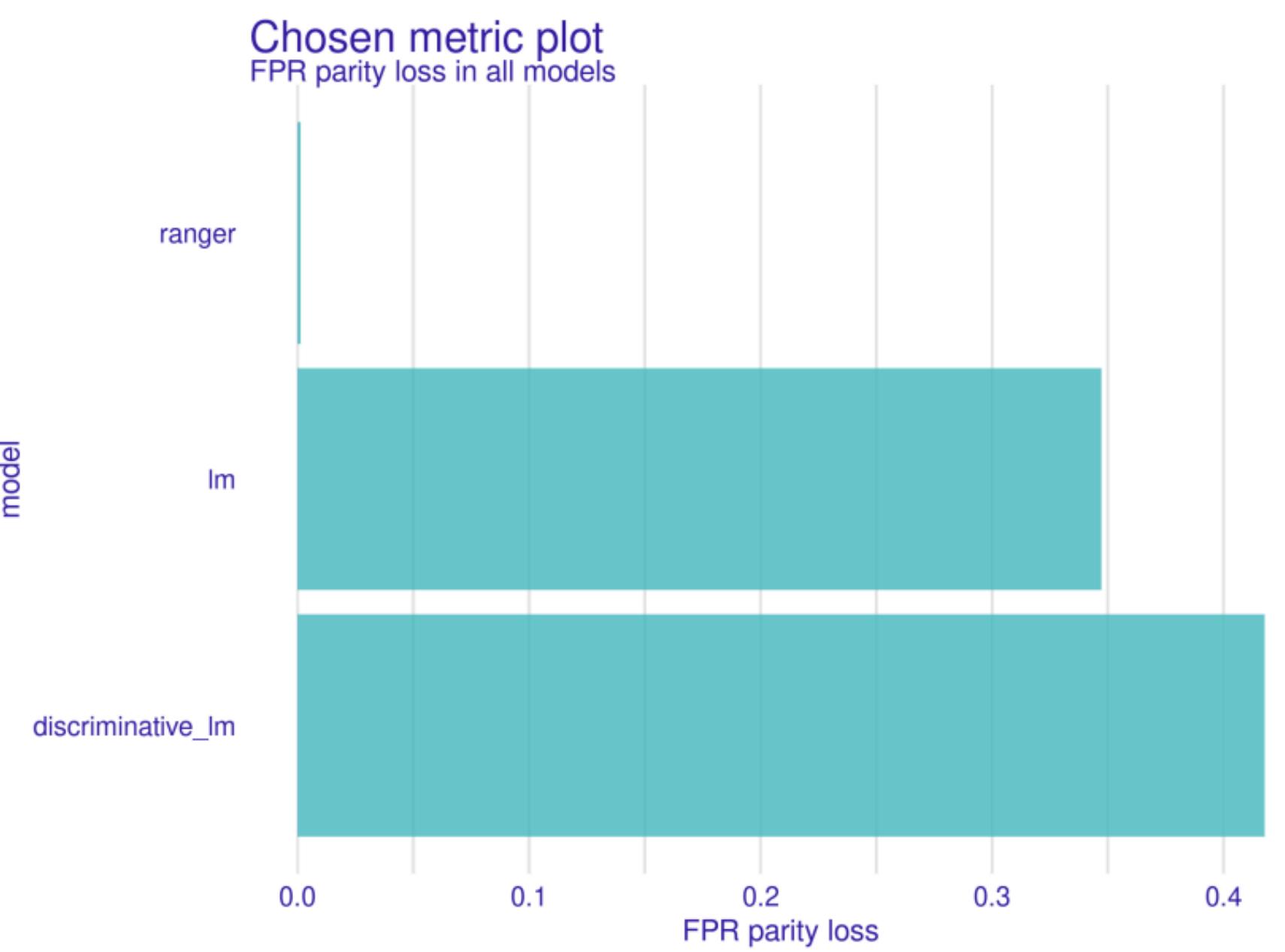
(b) fairness\_radar



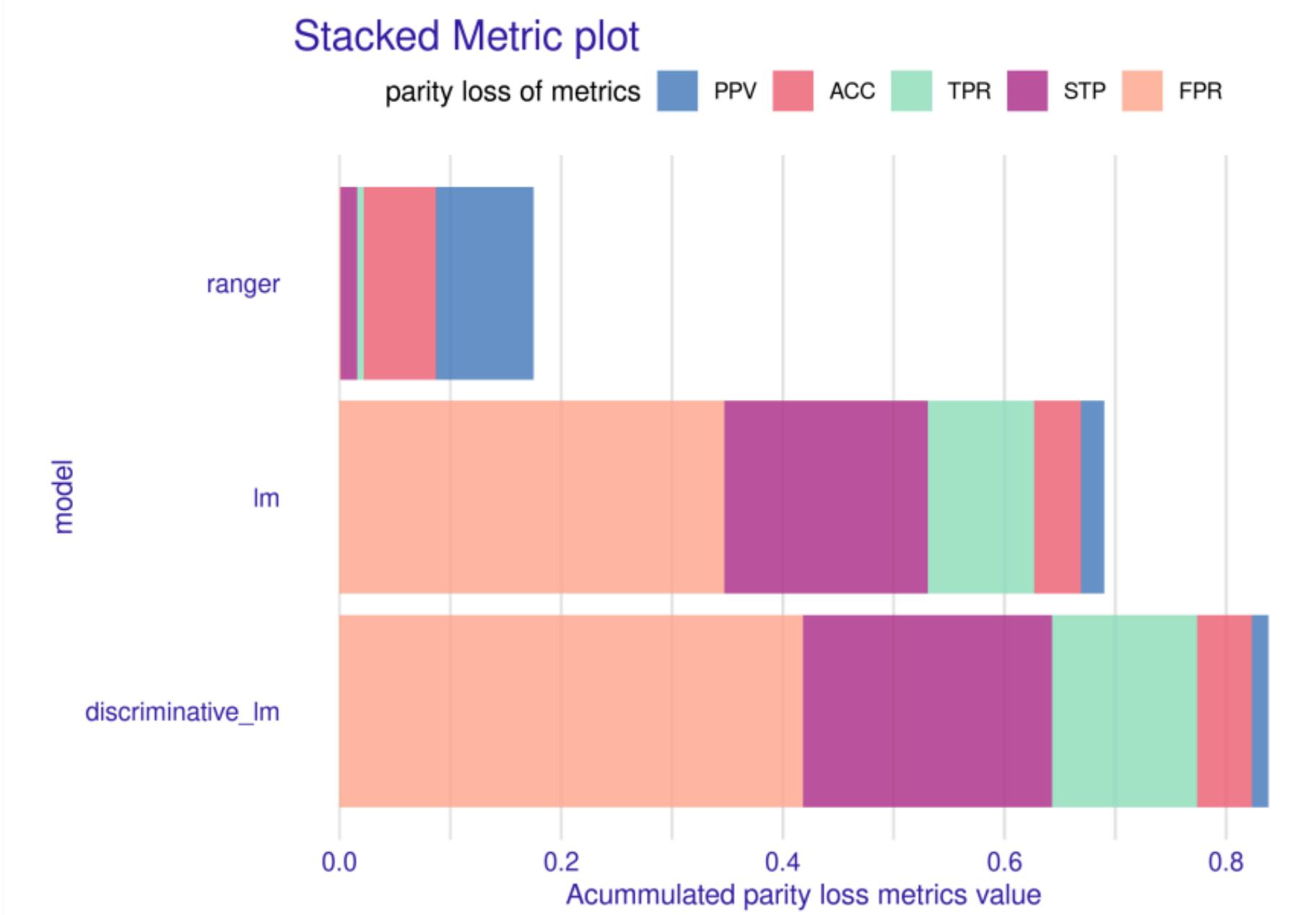
(c) fairness\_heatmap



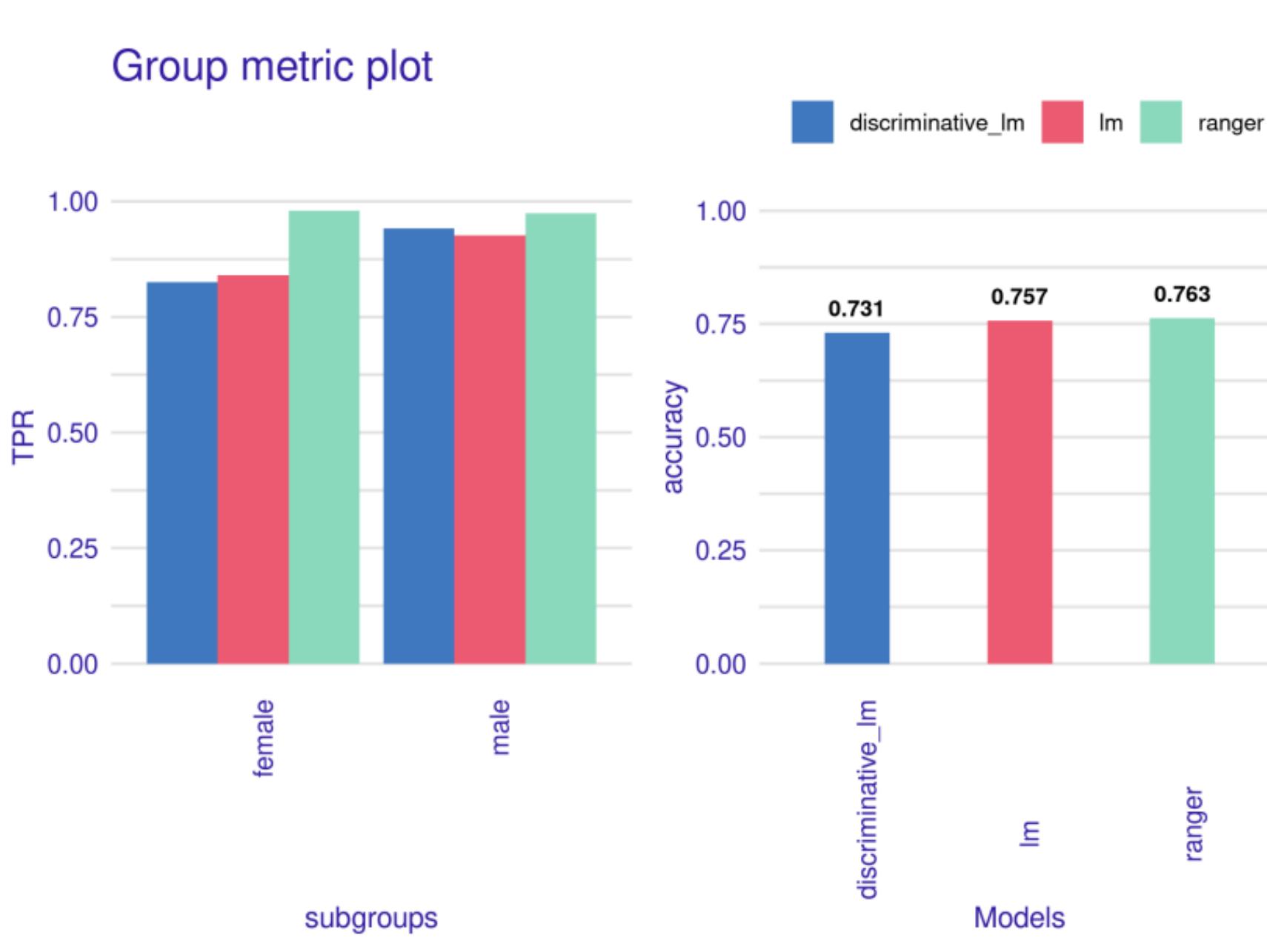
(d) fairness\_pca



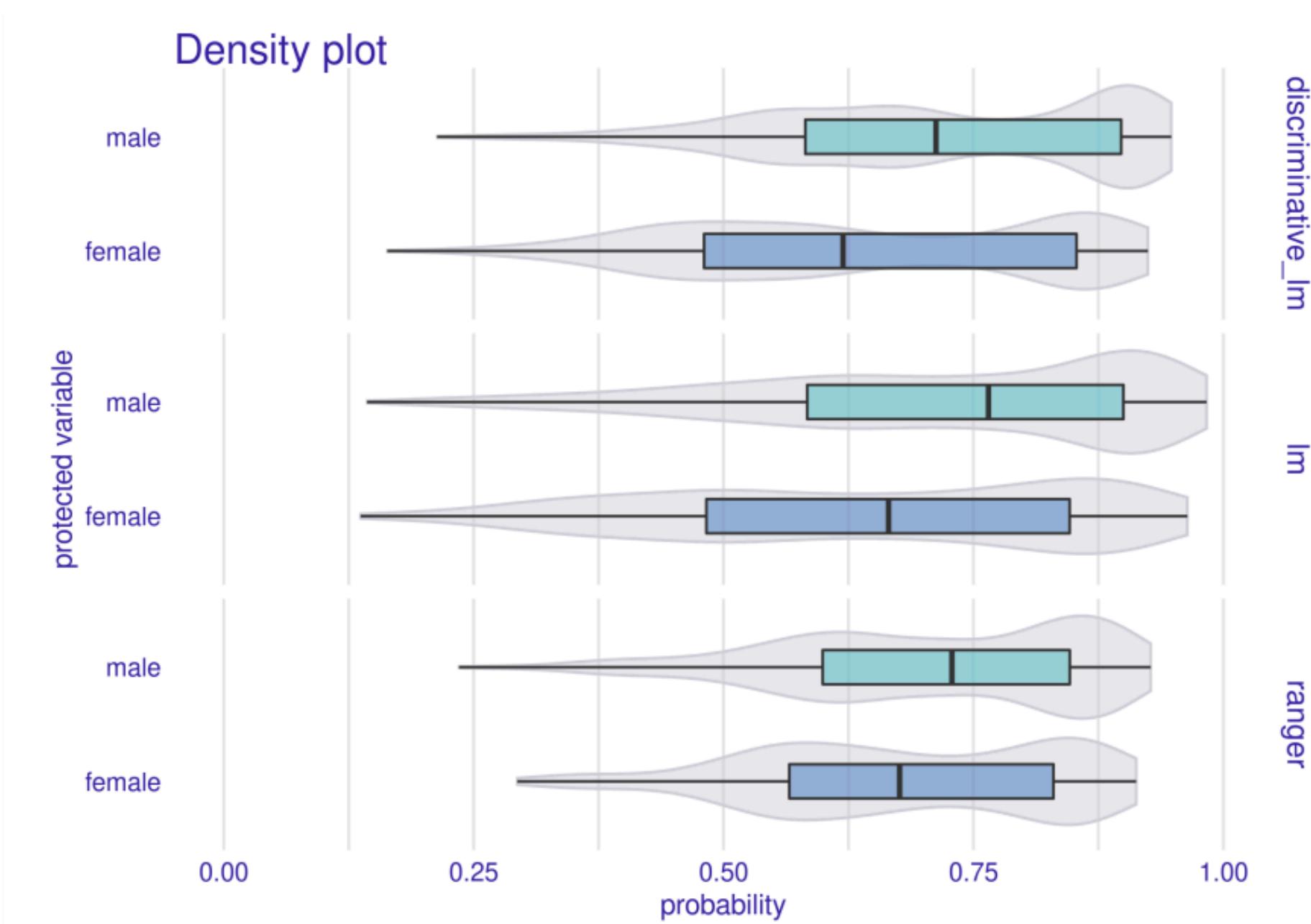
(e) choose\_metric



(f) stack\_metrics



(g) group\_metric



(h) plot\_density

## Performance and fairness plot



## AI Fairness 360 - Demo



Data    Check    Mitigate    Compare

### 1. Choose sample data set

Bias occurs in data used to train a model. We have provided three sample datasets that you can use to explore bias checking and mitigation. Each dataset contains attributes that should be protected to avoid bias.

#### Compas (ProPublica recidivism)

Predict a criminal defendant's likelihood of reoffending.

Protected Attributes:

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

[Learn more](#)

#### German credit scoring

Predict an individual's credit risk.

Protected Attributes:

- **Sex**, privileged: **Male**, unprivileged: **Female**
- **Age**, privileged: **Old**, unprivileged: **Young**

[Learn more](#)

#### Adult census income

Predict whether income exceeds \$50K/yr based on census data.

Protected Attributes:

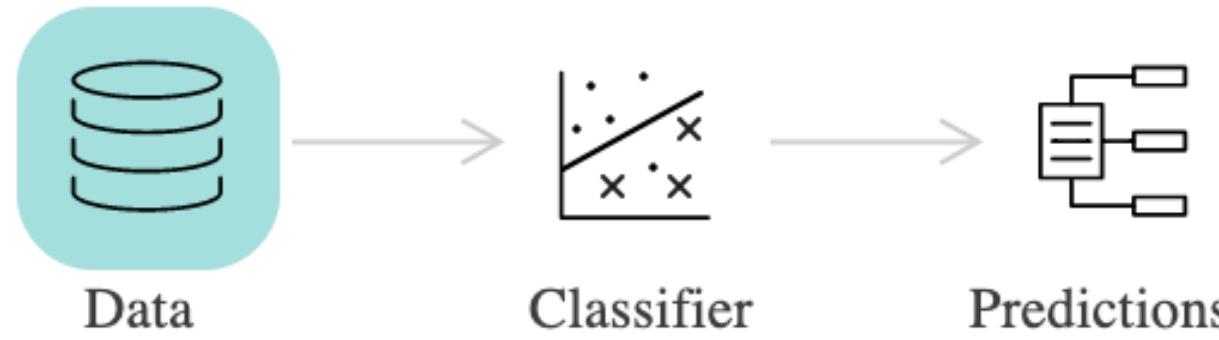
- **Race**, privileged: **White**, unprivileged: **Non-white**
- **Sex**, privileged: **Male**, unprivileged: **Female**

[Learn more](#)

**<http://aif360.mybluemix.net/data>**

## ○ Optimized Pre-Processing

Learns a probabilistic transformation that can modify the features and the labels in the training data.



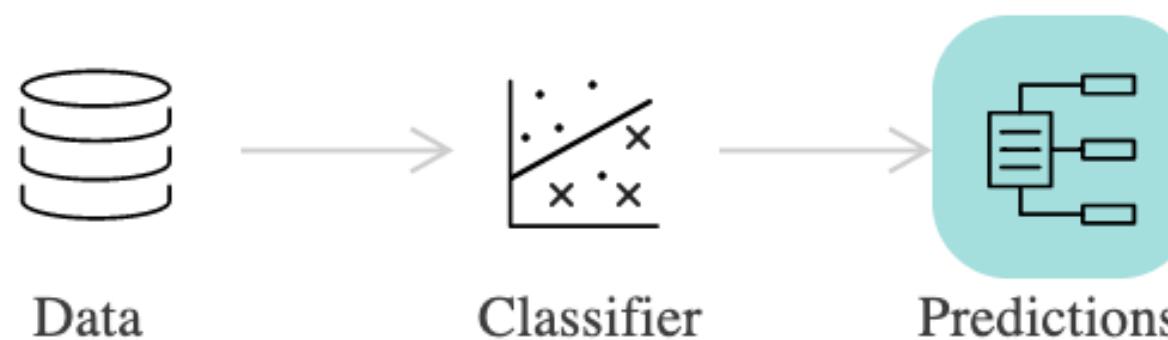
## ○ Adversarial Debiasing

Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.



## ○ Reject Option Based Classification

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



**<http://aif360.mybluemix.net/data>**

## 4. Compare original vs. mitigated results

Dataset: German credit scoring

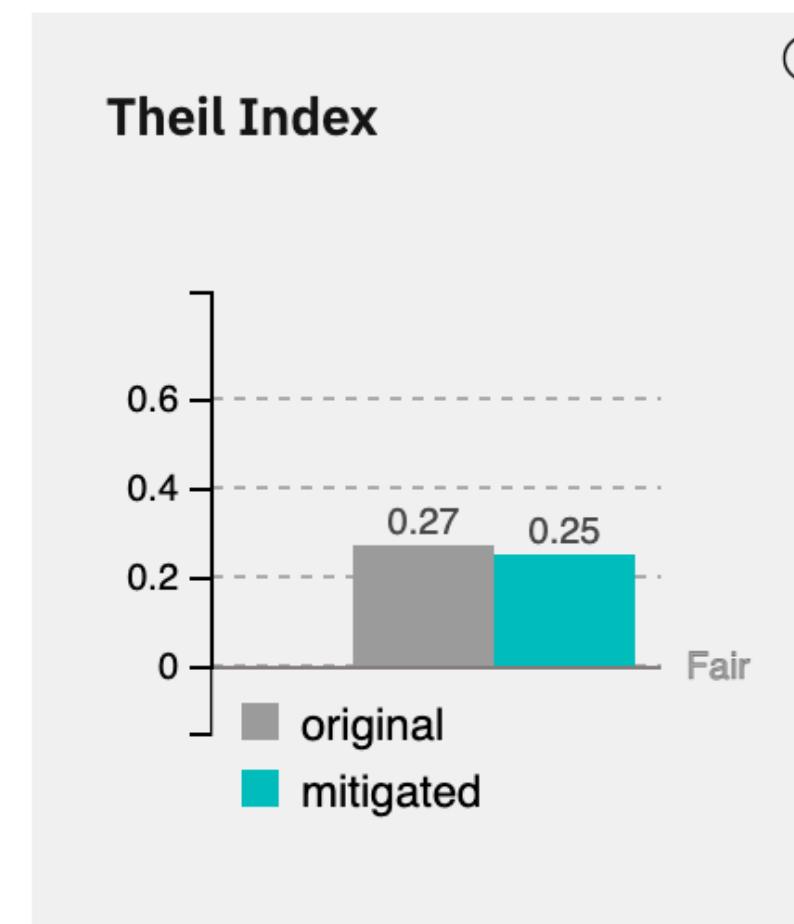
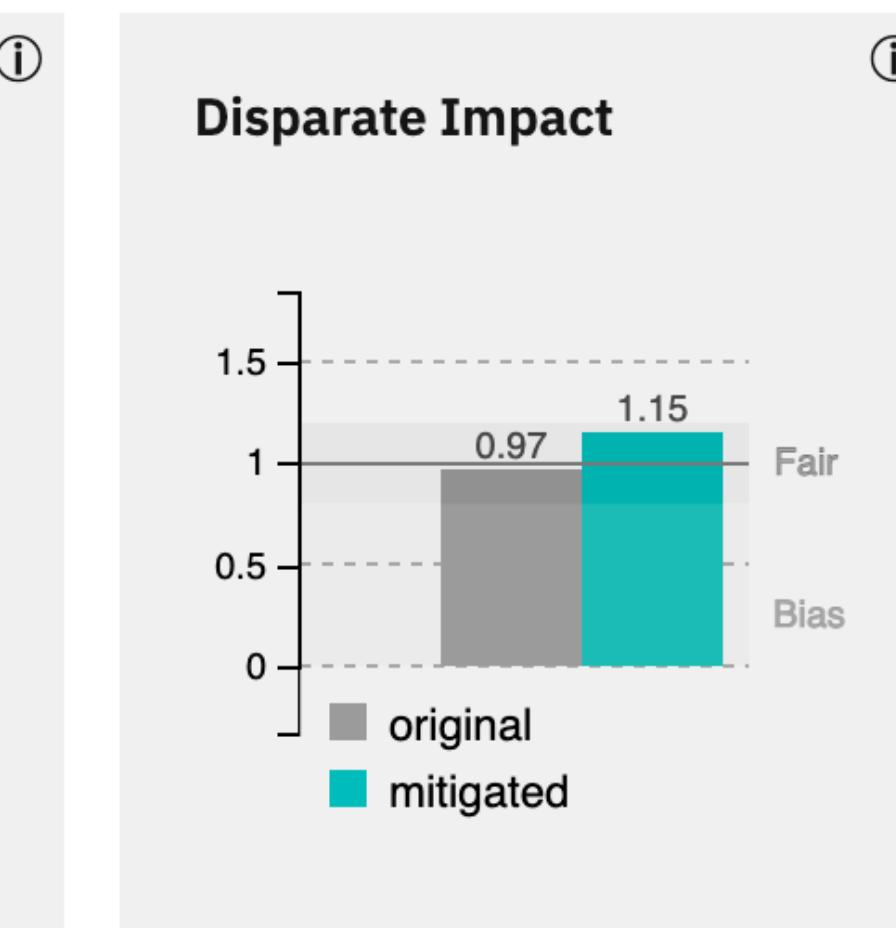
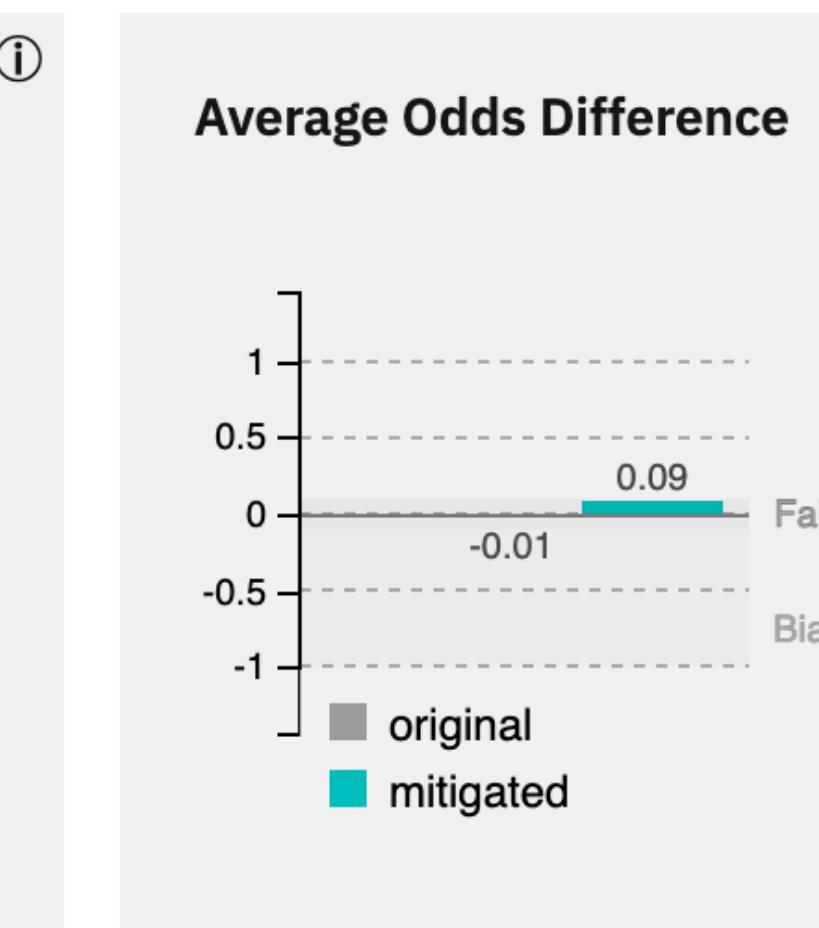
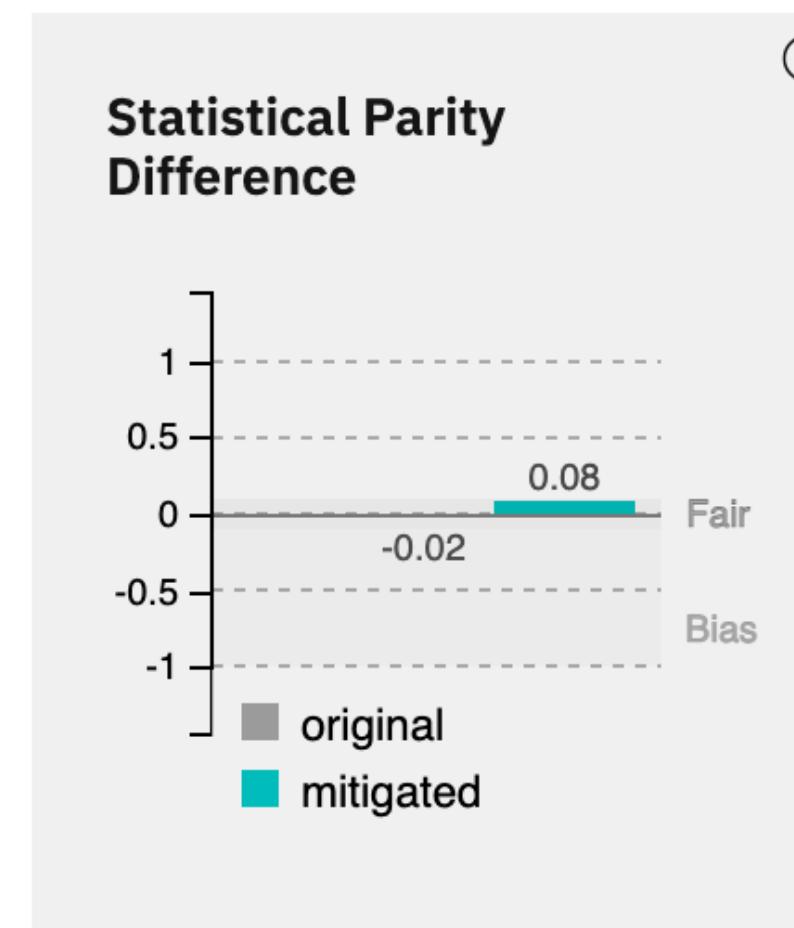
Mitigation: **Reject Option Based Classification algorithm applied**

### Protected Attribute: Sex

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

Accuracy after mitigation changed from 75% to 77%

Bias against unprivileged group unchanged after mitigation (0 of 5 metrics indicate bias)



# Questions? Comments?

Przemysław Biecek

<https://www.linkedin.com/in/pbiecek/>

Special thanks to MI2 DataLab:

Hubert Baniecki (modelStudio)  
Ewa Baranowska (drifter)  
Alicja Gosiewska (auditor)  
Aleksandra Grudziąż (survxai)  
Adam Izdebski (describe)  
Ewelina Karbowiak (ElX)  
Marcin Kosiński (archivist)  
Ania Kozak (vivo)  
Michał Kuźba (xaibot)  
Szymon Maksymiuk (DALEXtra)  
Magda Młynarczyk (cr17)  
Aleksandra Paluszyńska (randomForestExplainer)  
Kasia Pękała (tripplot)  
Piotr Piątyszek (Arena)  
Hanna Piotrowska (DrWhy theme)  
Adam Rydlewski (xai2cloud)  
Agnieszka Sitko (factorMerger)  
Jakub Wiśniewski (fairModels)  
... and others from MI2DataLab