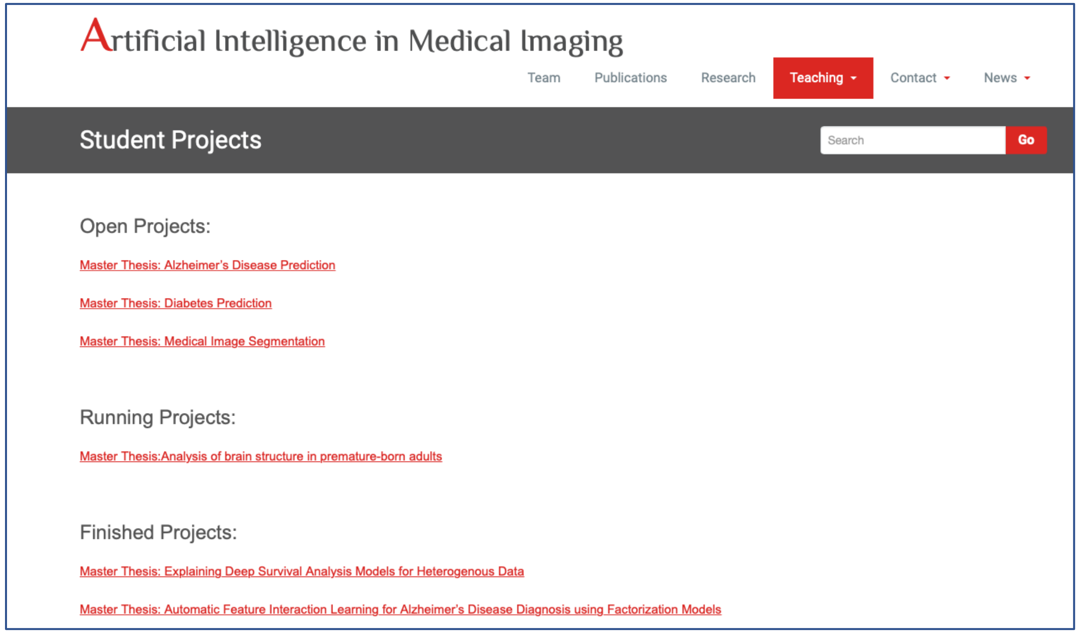
AI in Medicine I

Trustworthy AI: Fairness & Bias

Prof. Christian Wachinger
Lab for AI in Medical Imaging (www.ai-med.de)
Institute of Radiology, Klinikum rechts der Isar, TUM
School of Computation, Information and Technology, TUM

Student projects in my group



ai-med.de/research/student-projects/

Agenda

1. Trustworthy AI
2. Examples of unfair AI
3. Bias & Fairness
4. Fairness criteria
5. Metrics & Mitigation

Learning Outcomes

- **Understand importance** of trustworthy AI
- Awareness of **sources of bias** in AI
- **Quantify fairness** of prediction outcome
- Know strategies to **overcome bias**

Trustworthy AI

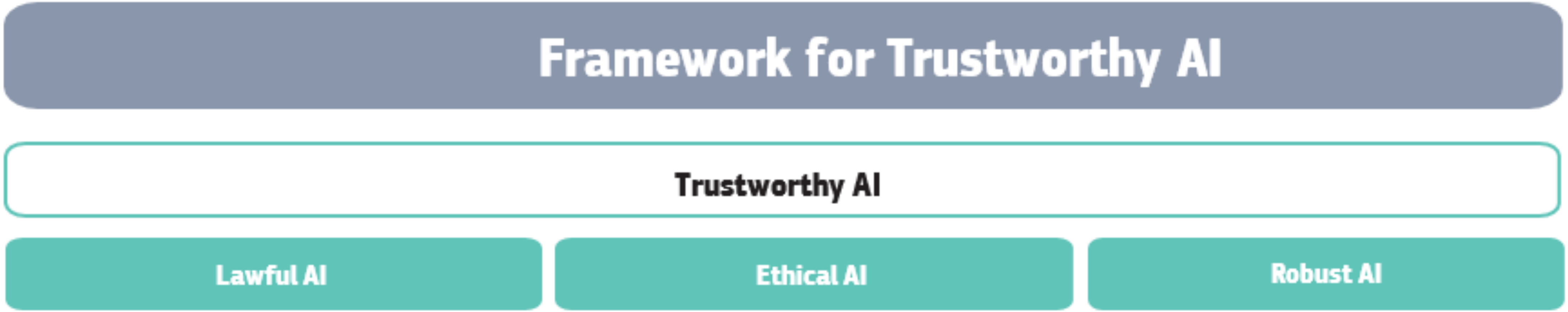
Why should we care about Trustworthy AI?

👍 Maximize the benefits of AI systems

👎 Minimize the risks

AI use needs trust of the society

Ethics guidelines for trustworthy AI
European Commission

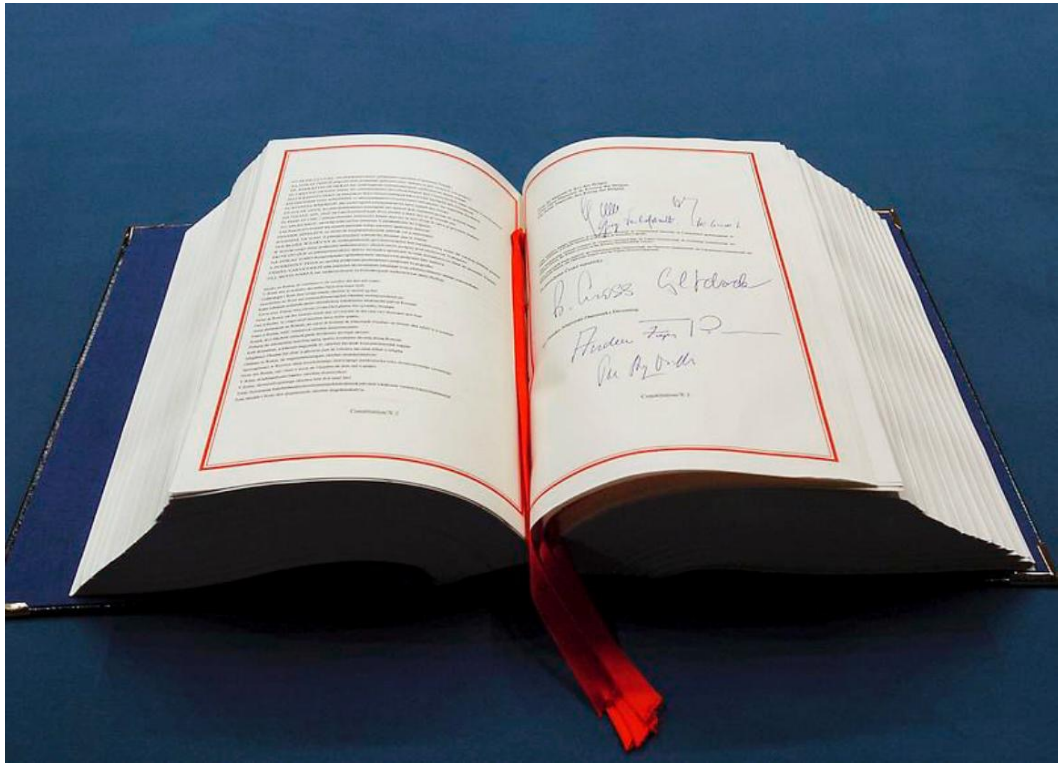


https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019

EU Charter of Fundamental Rights

… brings together the most important personal freedoms and rights enjoyed by citizens of the EU

- Dignity
- Freedoms
- Equality
- Solidarity
- Citizen’s rights
- Justice



Ethics guidelines for trustworthy AI
European Commission

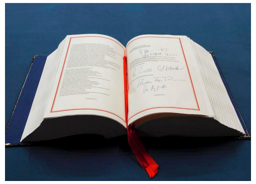
Framework for Trustworthy AI

Trustworthy AI

Lawful AI	Ethical AI	Robust AI

Foundations of Trustworthy AI → **4 Ethical Principles** →

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability



Adhere to ethical principles based on fundamental rights

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019

Ethics guidelines for trustworthy AI
European Commission

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability

Humans must be able to keep self-determination.
AI systems should not deceive, or manipulate humans.

AI systems should not harm humans.
Protection of human dignity as well as mental and physical integrity.

Processes need to be transparent, and decisions explainable

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019

Fairness

Ensuring equal and just distribution of both benefits and costs.

Ensuring that individuals and groups are free from unfair bias.
Unfair bias could lead to
- Marginalization of vulnerable groups
- Amplification of prejudice and discrimination.

Stated principles remain **abstract ethical prescriptions**. AI practitioners can hence not be expected to find the right solution based on the principles above, yet they should **approach ethical dilemmas via reasoned, evidence-based reflection** rather than intuition.

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019

Discussion among students

Do you know UNFAIR or BIASED AI systems?

Consider general scenarios and medical ones.

5 minutes

Examples of unfair AI

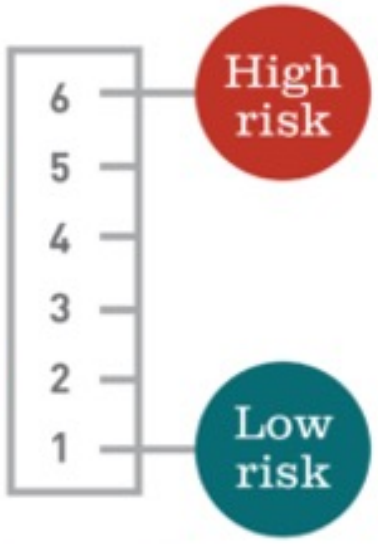
AI systems in the courtroom

Relased to await trail

Detained in jail to await trail

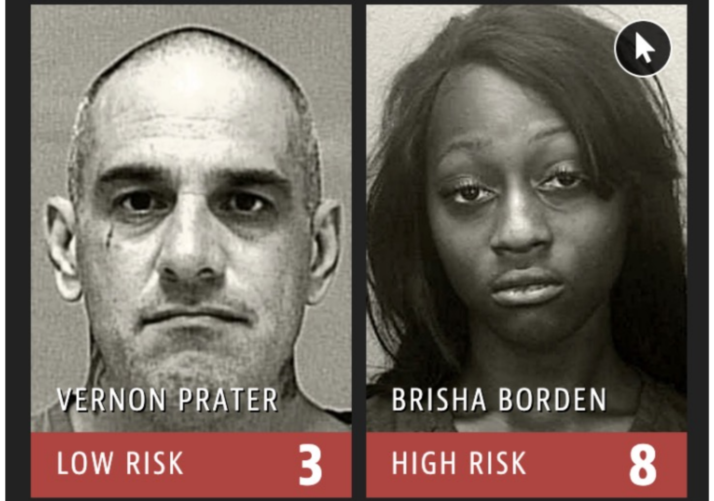
AI support: Predict risk for

- Future crimes
- Failure to appear in court



Controversy over the COMPAS Score (Propublica)

Two Petty Theft Arrests

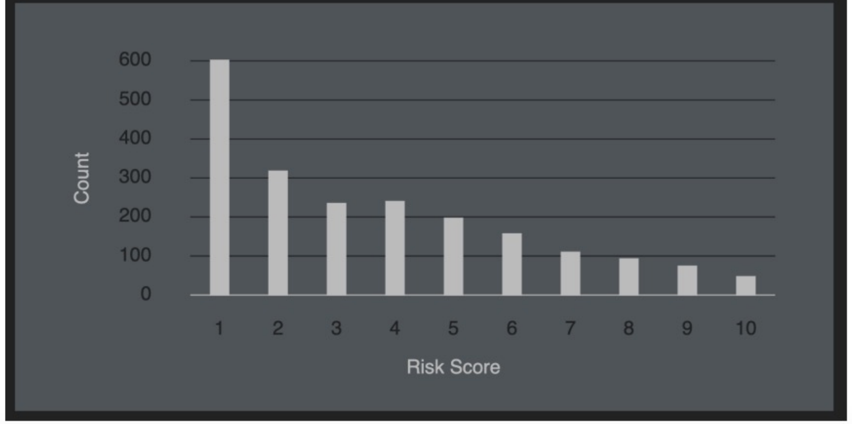


Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

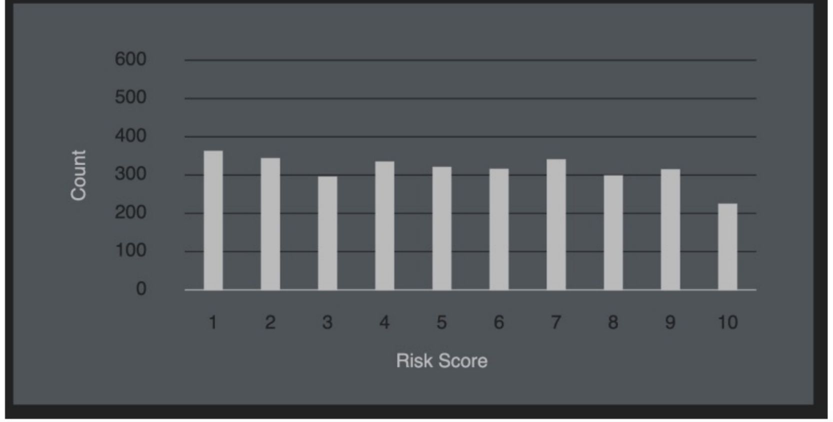
Two Drug Possession Arrests

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

White Defendants' Risk Scores



Black Defendants' Risk Scores

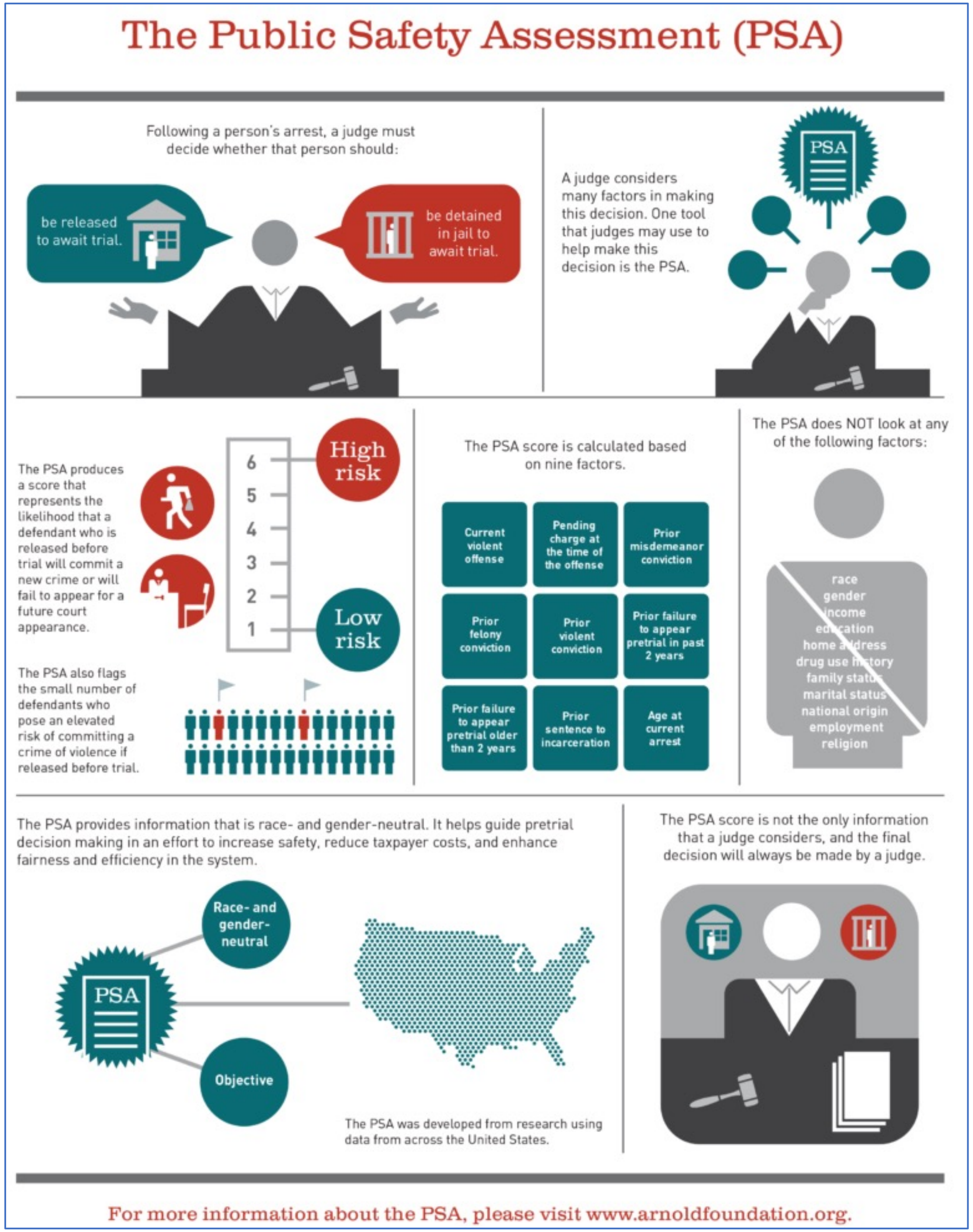


https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

AI systems in the courtroom

“... the machine learning systems used to calculate these risk scores throughout the criminal justice system, have been shown to hold **severe racial biases**, scoring people of color more likely to commit future crimes.”

https://pretrialrisk.com/the-psa-is-spreading-and-fast/



The Public Safety Assessment (PSA)
Following a person's arrest, a judge must decide whether that person should:
be released to await trial.
be detained in jail to await trial.
A judge considers many factors in making this decision. One tool that judges may use to help make this decision is the PSA.
PSA
The PSA produces a score that represents the likelihood that a defendant who is released before trial will commit a new crime or will fail to appear for a future court appearance.
6
5
4
3
2
1
High risk
Low risk
The PSA also flags the small number of defendants who pose an elevated risk of committing a crime of violence if released before trial.
The PSA score is calculated based on nine factors.
Current violent offense
Pending charge at the time of the offense
Prior misdemeanor conviction
Prior felony conviction
Prior violent conviction
Prior failure to appear pretrial in past 2 years
Prior failure to appear pretrial older than 2 years
Prior sentence to incarceration
Age at current arrest
The PSA does NOT look at any of the following factors:
race
gender
income
education
home address
drug use history
family status
marital status
national origin
employment
religion
The PSA provides information that is race- and gender-neutral. It helps guide pretrial decision making in an effort to increase safety, reduce taxpayer costs, and enhance fairness and efficiency in the system.
Race- and gender-neutral
PSA
Objective
The PSA was developed from research using data from across the United States.
The PSA score is not the only information that a judge considers, and the final decision will always be made by a judge.
For more information about the PSA, please visit www.arnoldfoundation.org.

The Intercept_

Illustration: The Intercep

BECOME A MEMBER

A BAIL REFORM TOOL INTENDED TO CURB MASS INCARCERATION HAS ONLY REPLICATED BIASES IN THE CRIMINAL JUSTICE SYSTEM

Kentucky was an early adopter of risk assessments in an effort to release more people without bail. But the algorithms are reproducing systemic inequities.

Bryce Covert
July 12 2020, 2:00 p.m.

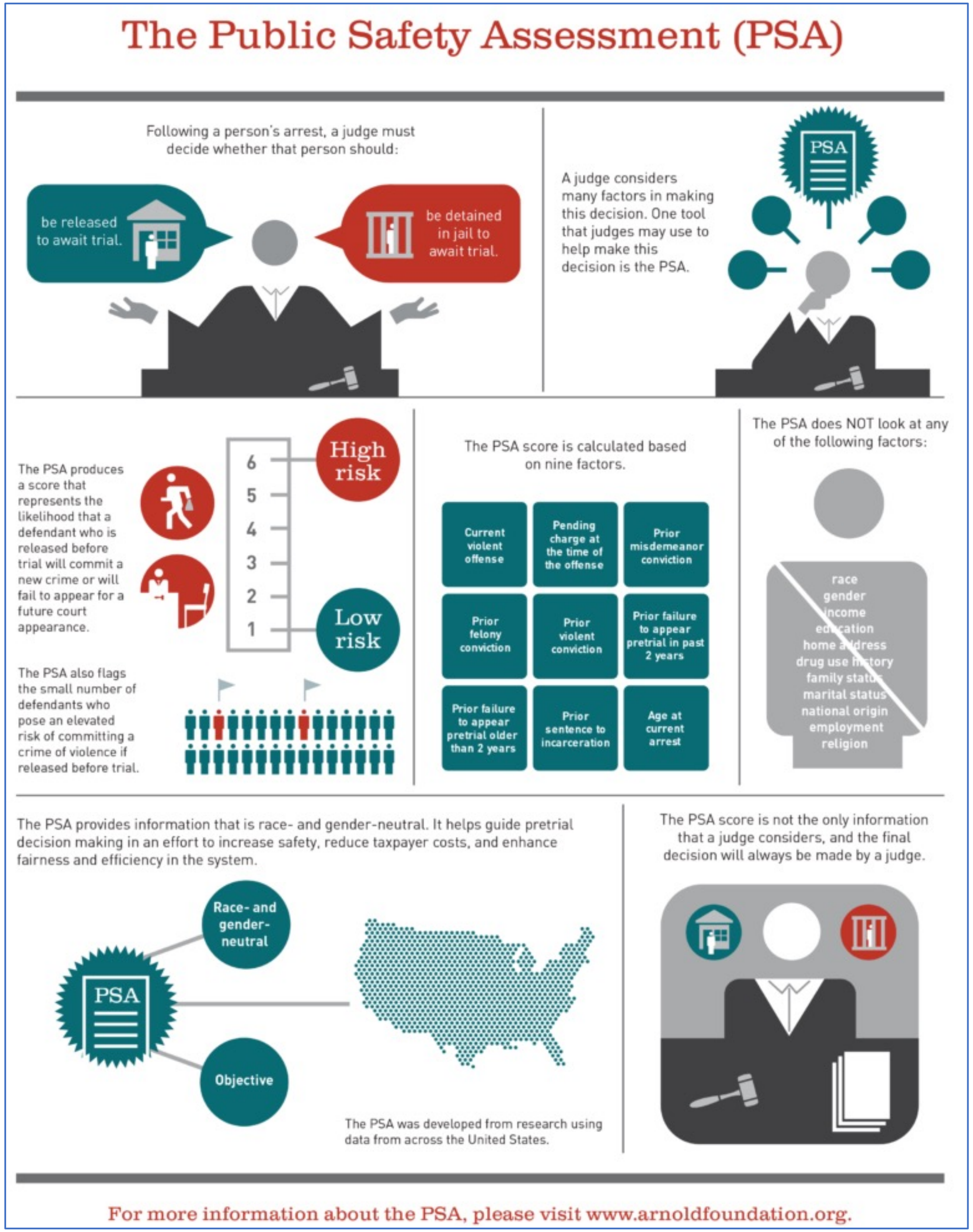
Wisconsin Watch

Recent stories | Major projects | Coverage areas | News414 | Audio stories | Store | Be your own watchdog

BEYOND BAIL | JUSTICE & SAFETY

Debate rages over whether pretrial risk assessments are racially biased

Some researchers say there is not enough evidence to know whether the tools will reduce or exacerbate discrimination that already exists in the justice system



"Prediction **looks to the past** to make guesses about the future. In a racially stratified world, any method of prediction will project the inequalities of the past into the future."

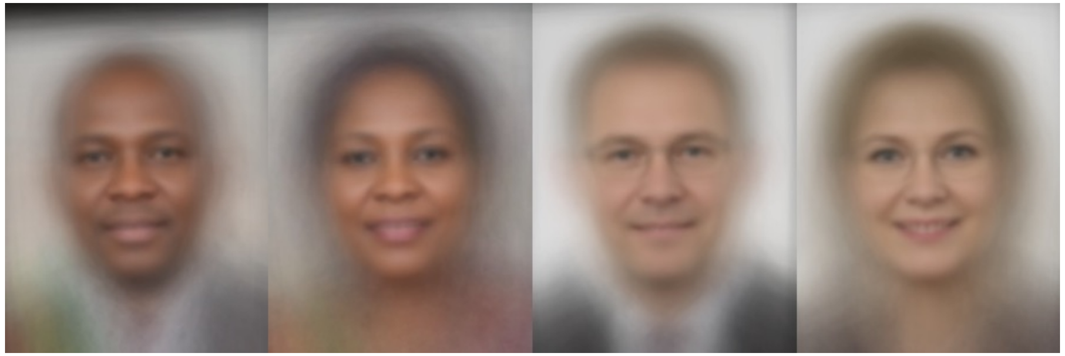
https://digitalcommons.law.uga.edu/fac_artchop/1293/

AI in recruiting

- Amazon hiring tool to mechanize search for top talent
- Found to be biased against women, penalizing candidates whose resumes included the word “women’s”.

https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf

Gender Shades Project



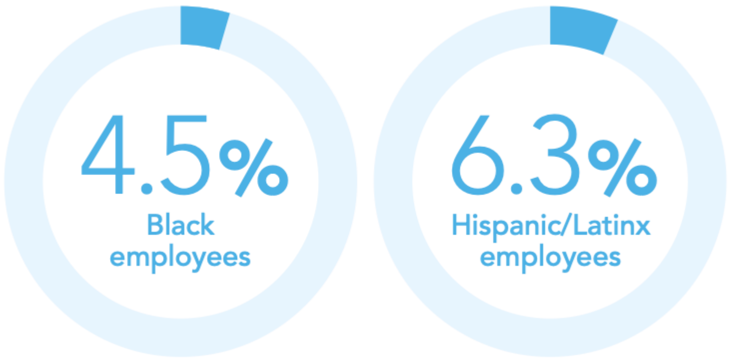
Lack of diversity in the training data

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%

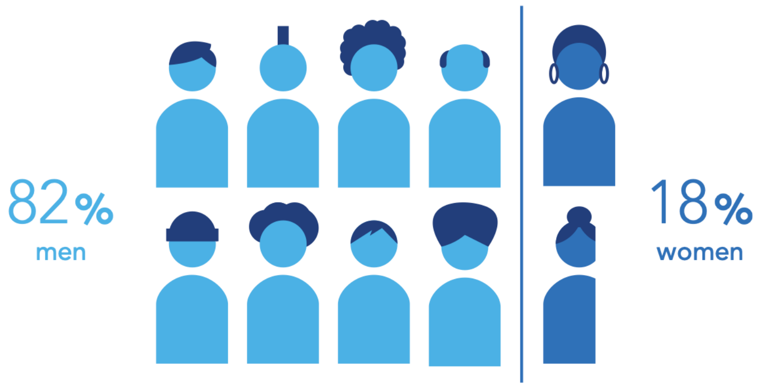
http://gendershades.org/, Buolamwini, Gebru, PMLR, 2018

Gender Shades Project

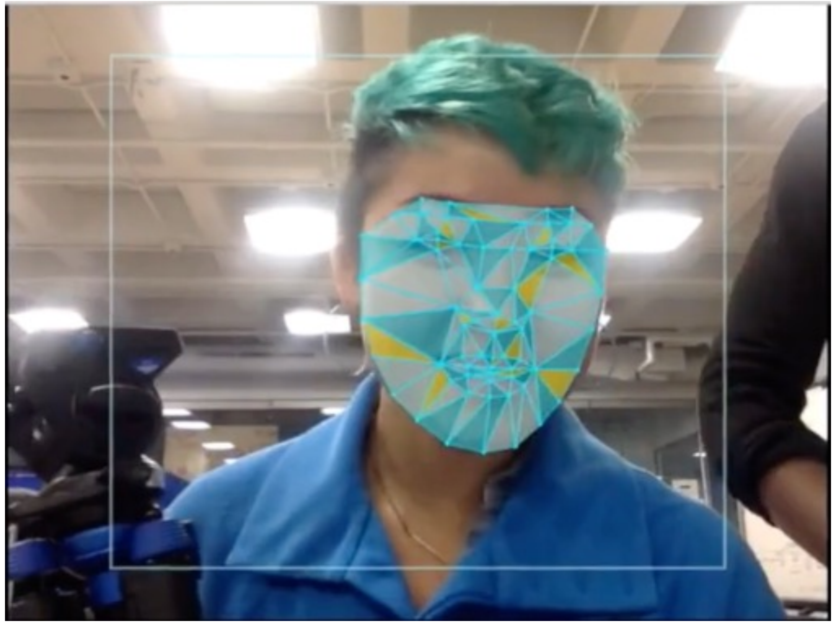
Employees at Microsoft

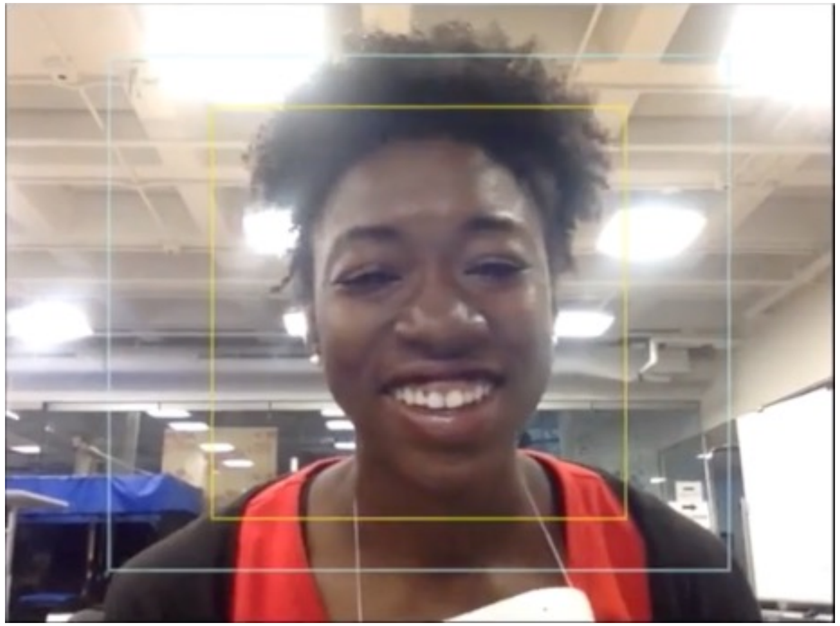


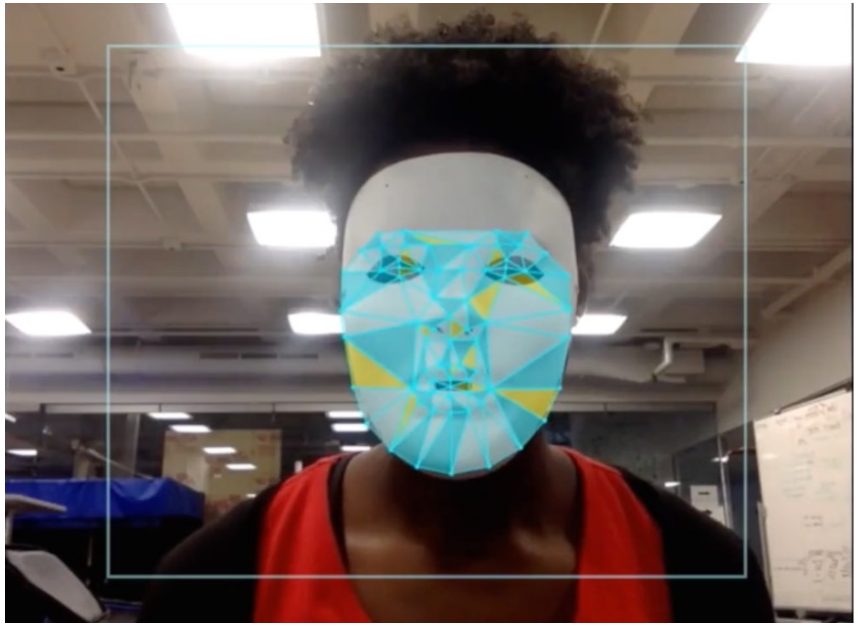
Researchers at leading AI conferences



The Coded Gaze: Unmasking Algorithmic Bias







https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

ImageNet et al. criticism

Large datasets are instrumental for DL success

Dataset	Number of images (in millions)	Number of categories (in thousands)	Number of consensual images
JFT-300M ([41])	300+	18	0
Open Images ([50])	9	20	0
Tiny-Images ([79])	79	76	0
Tencent-ML ([89])	18	11	0
ImageNet-(21K,11k[1],1k) ([70])	(14, 12, 1)	(22, 11, 1)	0
Places ([93])	11	0.4	0

Grabbing images from the internet **without consensus** esp. children

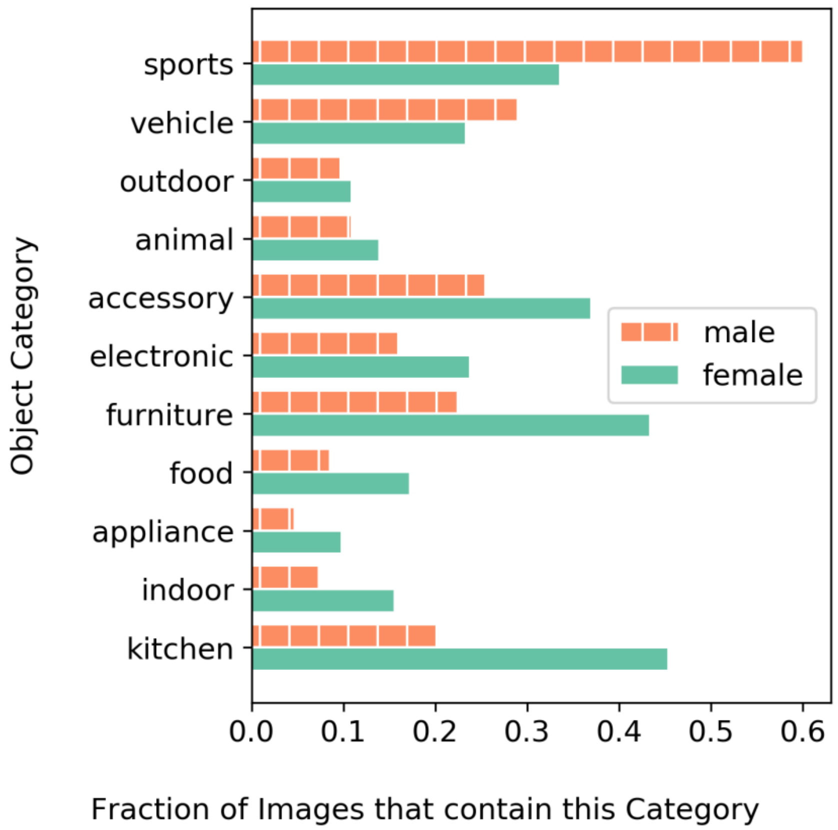
Categorize people based on appearance

Dataset audit cards for curators: publish the goals, curation procedures, known shortcomings and caveats

Birhane, Prabhu, Large image datasets: A pyrrhic win for computer vision? *WACV, 2021*

Measuring Bias in Visual Datasets

In which image contexts appear men and women?



https://arxiv.org/pdf/2004.07999.pdf

AI in health care

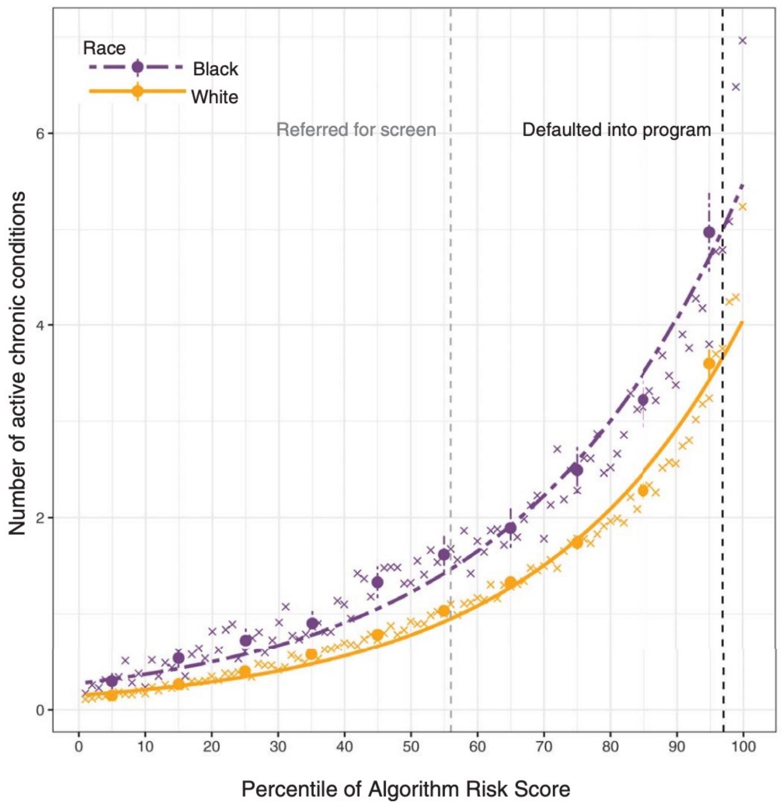
Task: Identify patients for high-risk care management
Algorithm: Predicts risk score for each patient
Policy: Patients above the 97th percentile are enrolled

Algorithm input: demographics (e.g., age, sex), insurance type, diagnosis and procedure codes, medications, and detailed costs. **Excludes race.**
Output: health costs (costs as a proxy for health needs)

Obermeyer, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 2019.

AI in health care

At the same level of algorithm-predicted risk,
Blacks have significantly more illness burden than Whites



Task: Identify patients for high-risk care management
Algorithm: Predicts risk score for each patient
Policy: Patients above the 97th percentile are enrolled

Algorithm input: demographics (e.g., age, sex), insurance type, diagnosis and procedure codes, medications, and detailed costs. **Excludes race.**
Output: health costs (costs as a proxy for health needs)

Black patients generate lesser medical expenses (unequal access to care)
=> algorithm learned that Black patients are healthier than equally sick White patients

Obermeyer, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 2019.

Bias in Medical Imaging

Skin lesions

AI skin cancer diagnoses risk being less accurate for dark skin – study

Research finds few image databases available to develop technology contain details on ethnicity or skin type

Nicola Davis *Science correspondent*

@NicolaKSDavis

Tue 9 Nov 2021 23.30 GMT

Studies suggest image recognition technology can classify skin cancers as successfully as humans. Posed by model. Photograph: ChesiireCat/Getty Images/iStockphoto

Latest

The Atlantic

HEALTH

AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind

Machine learning has the potential to save thousands of people from skin cancer each year—while putting others at greater risk.

By Angela Lashbrook

MOTHERBOARD

TECH BY VICE

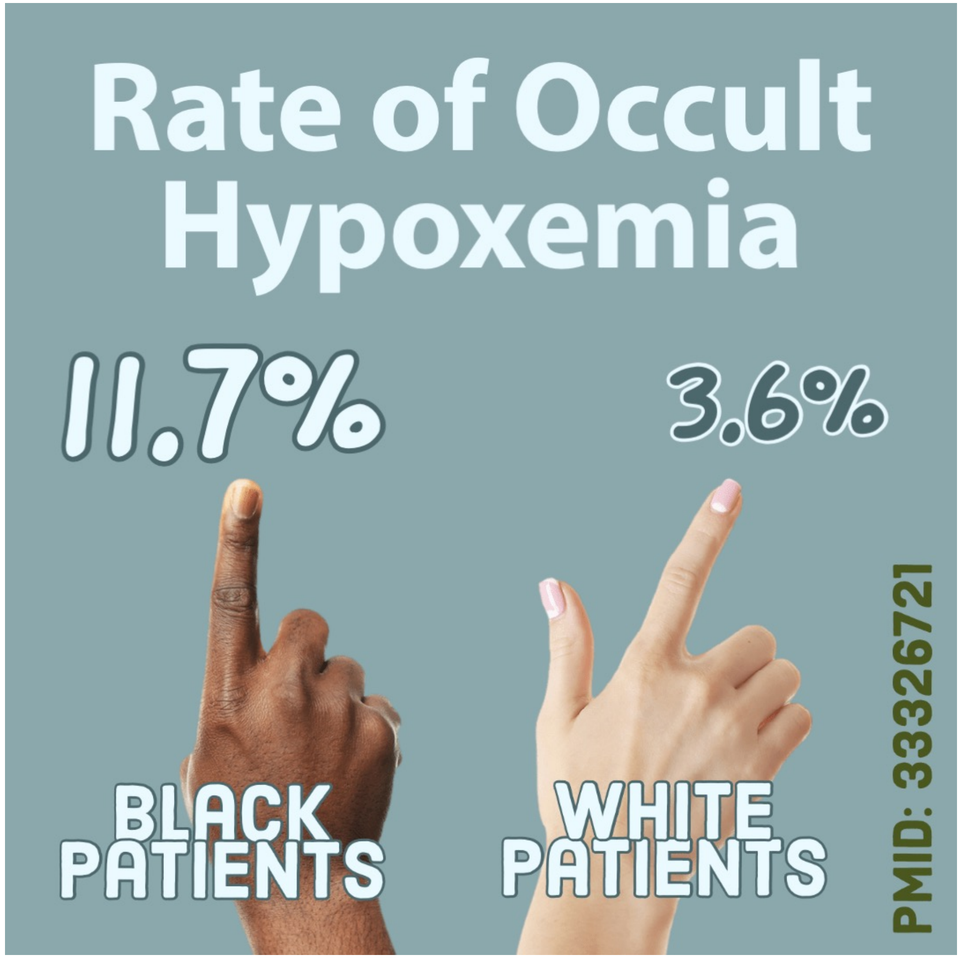
Google's New Dermatology App Wasn't Designed for People With Darker Skin

The company trained the system to recognize different skin conditions. But like Google itself, the app's data has a diversity problem.

TF By Todd Feathers

May 20, 2021, 3:40pm Share Tweet Snap

Pulse Oximeters: Falsely elevated oxygen saturation in dark-skinned persons



Occult hypoxemia: falsely elevated oxygen saturation when measured with pulse oximetry, masking hypoxemia

https://jamanetwork.com/journals/jama/fullarticle/2800468

Impact of user biases while using AI tools

npj | digital medicine

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › npj digital medicine › perspectives › article

Perspective | Open Access | Published: 28 December 2022

AI in the hands of imperfect users

Kristin M. Kostick-Quenet & Sara Gerke

npj Digital Medicine **5**, Article number: 197 (2022) | Cite this article

2872 Accesses | **34** Altmetric | Metrics

https://www.nature.com/articles/s41746-022-00737-z

https://www.healthcareitnews.com/news/implementation-best-practices-dealing-complexity-ai

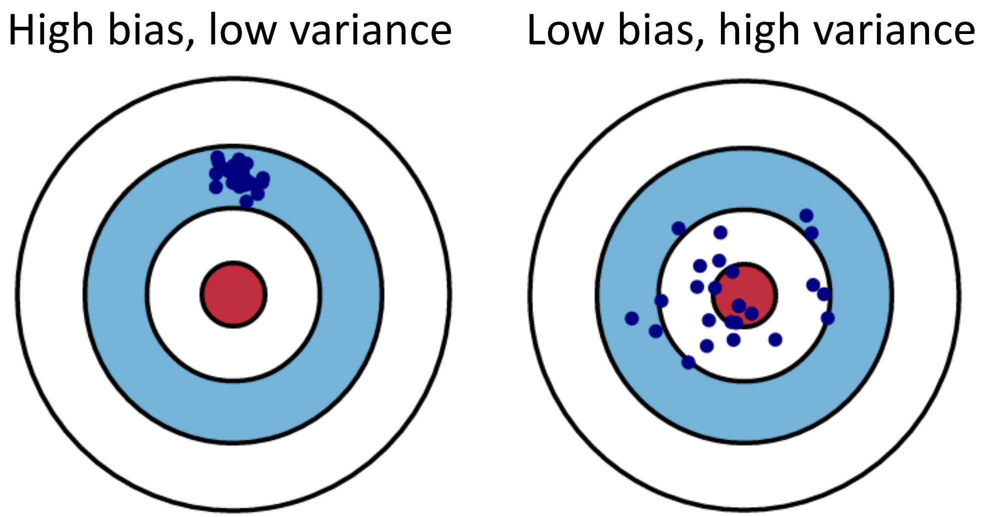
BIAS & FAIRNESS

Bias

Biases are **cognitive shortcuts** that can result in judgments which lead to discriminatory practices.

Bias in statistics:
Bias is the tendency of a statistic to **systematically** overestimate or underestimate a population parameter

Bias-Variance tradeoff in machine learning



Fairness == absence of bias?

WA is fairness the absence of bias?

Fairness and bias are related concepts, but they are not the same thing. Fairness generally refers to the idea that all individuals should be treated equally, regardless of their characteristics such as race, gender, or age. Bias, on the other hand, refers to the systematic and unjustified difference in treatment of individuals based on those same characteristics.

In the context of machine learning, fairness can be thought of as the absence of bias in the model's predictions or decisions. This means that the model should not make predictions or decisions that are systematically different for certain groups of individuals based on their protected characteristics. However, it's important to keep in mind that achieving fairness in machine learning models is a complex task that considers multiple aspects and trade offs, and it's not just a matter of eliminating bias.

e a

gh variance

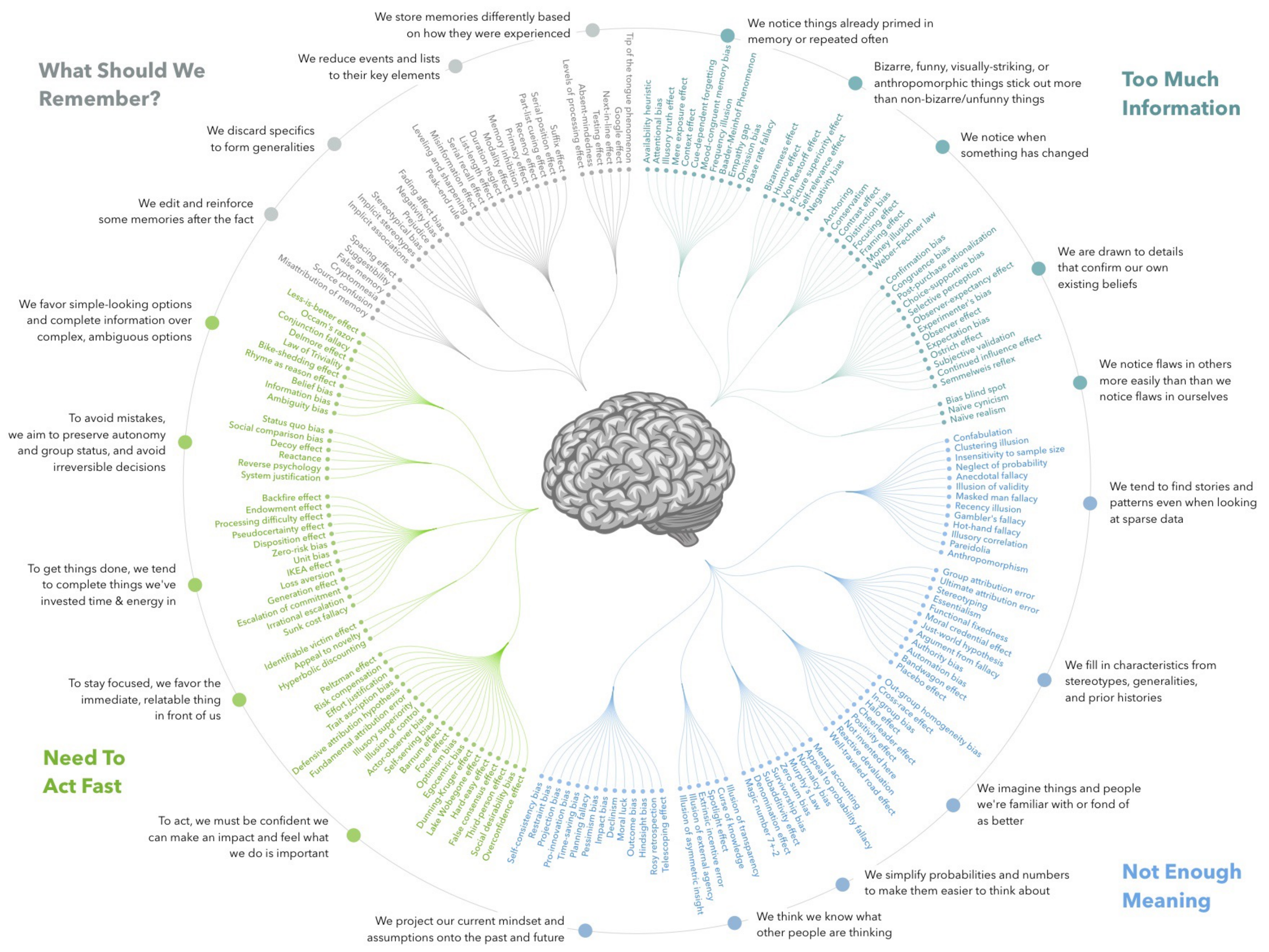
Fairness == absence of bias?

Common Biases

- **Selection bias** (sampling bias): selection of individuals such that the sample is NOT representative of the population intended to be analyzed

- **Confounding bias**: more on this in AIM II

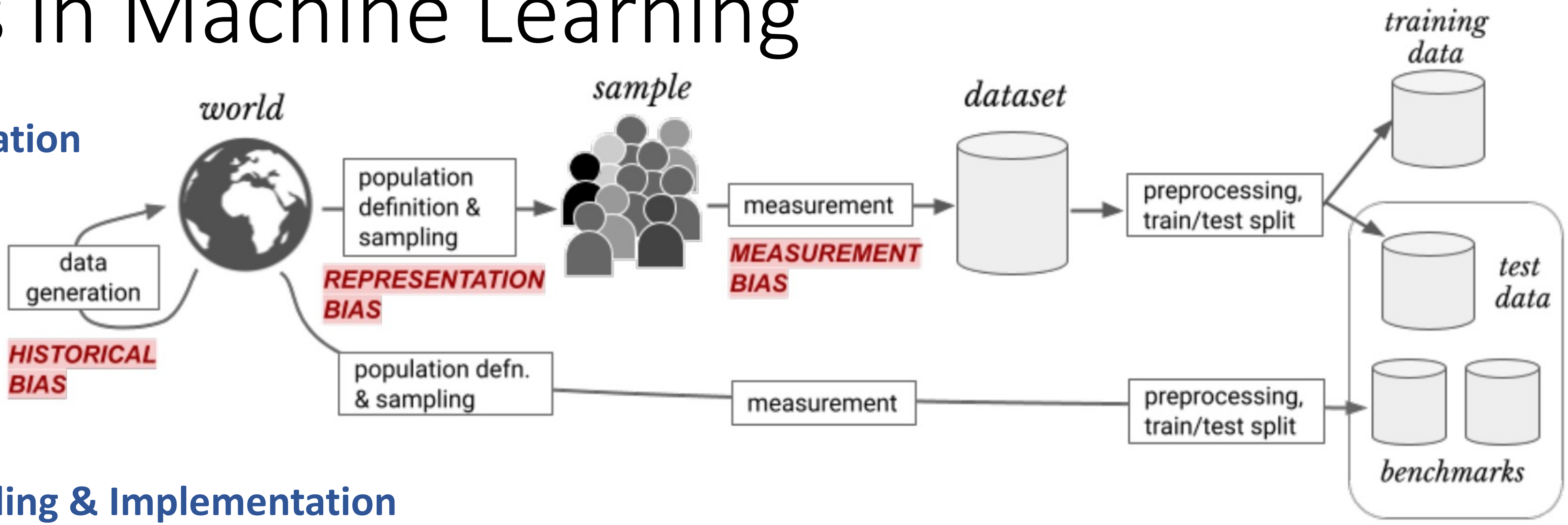
COGNITIVE BIAS CODEX



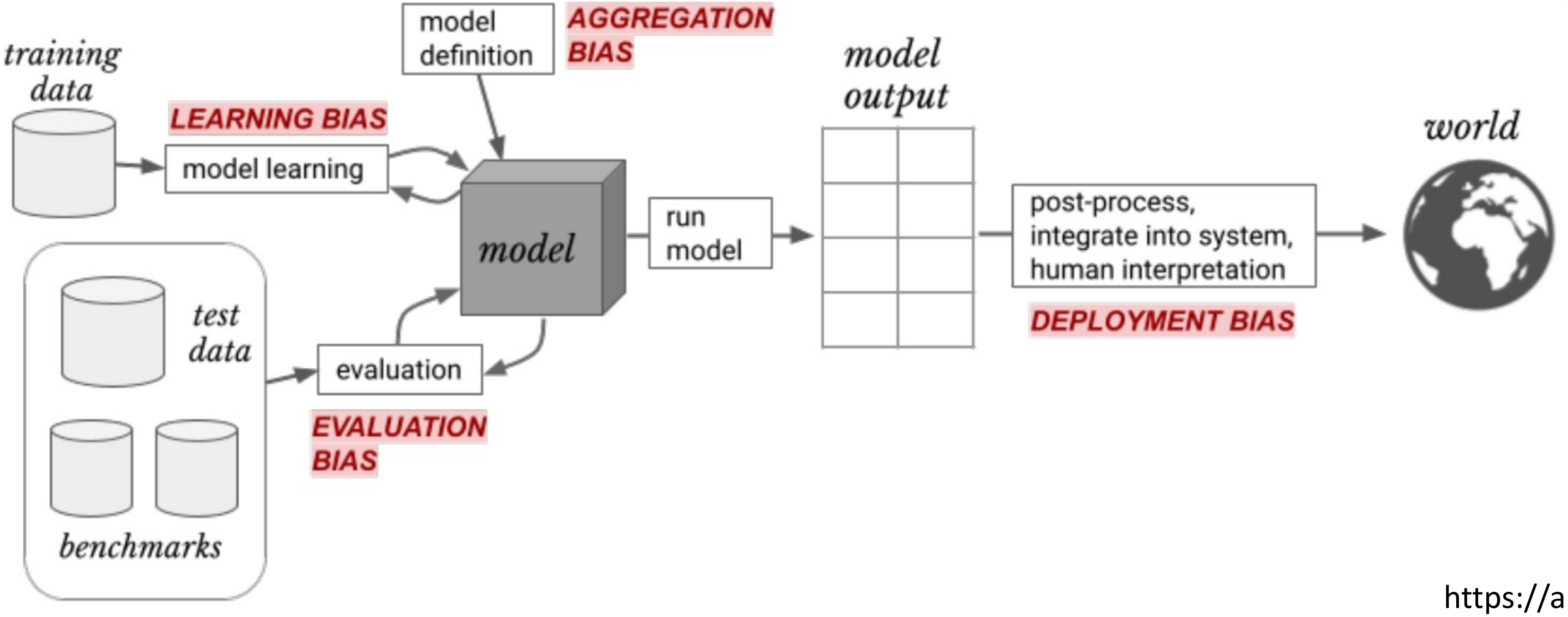
DESIGNHACKS.CO · CATEGORIZATION BY BUSTER BENSON · ALGORITHMIC DESIGN BY JOHN MANOOGIAN III (JM3) · DATA BY WIKIPEDIA

Biases in Machine Learning

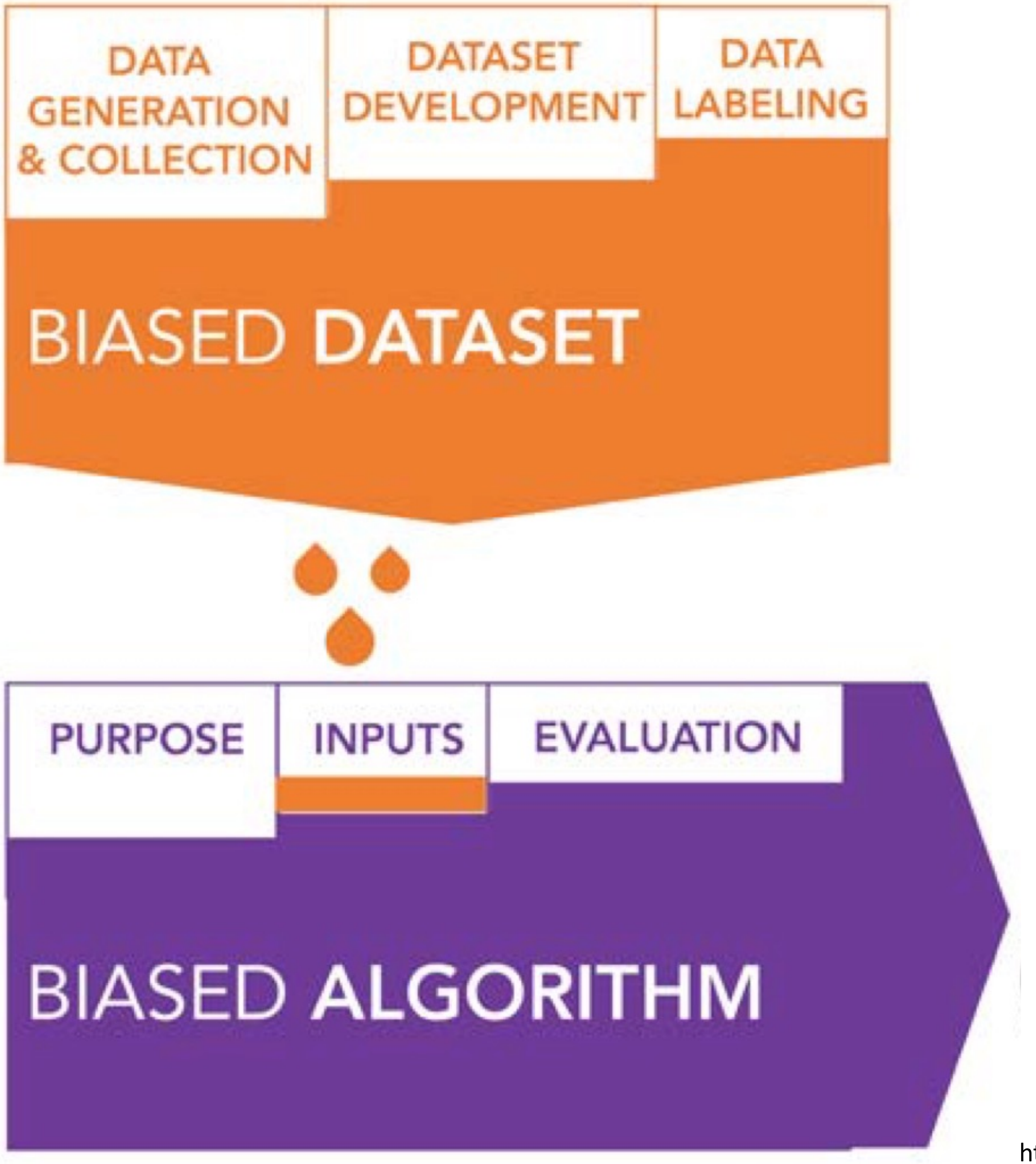
Data Generation



Model Building & Implementation



https://arxiv.org/pdf/1901.10002.pdf

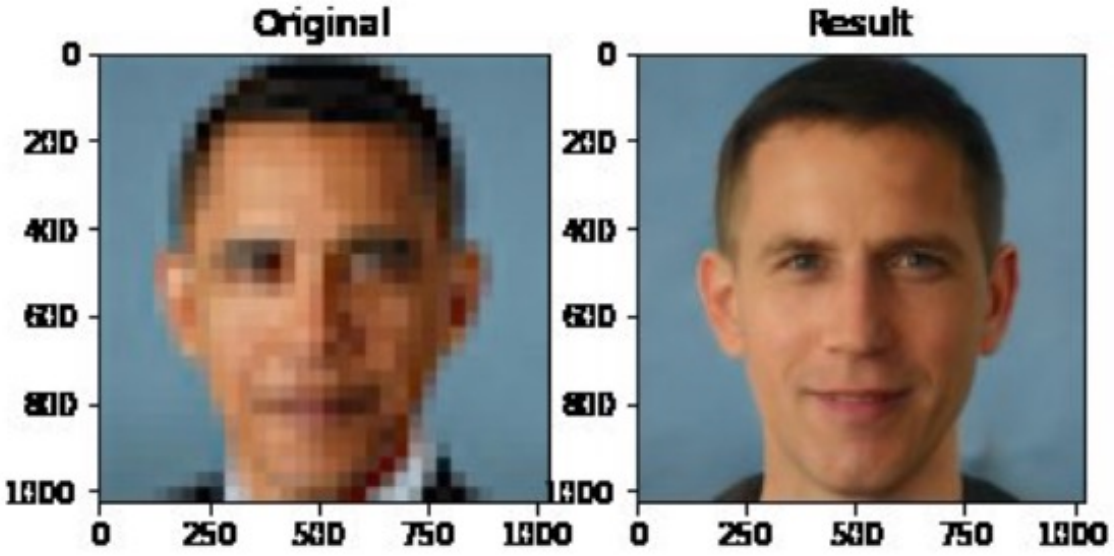


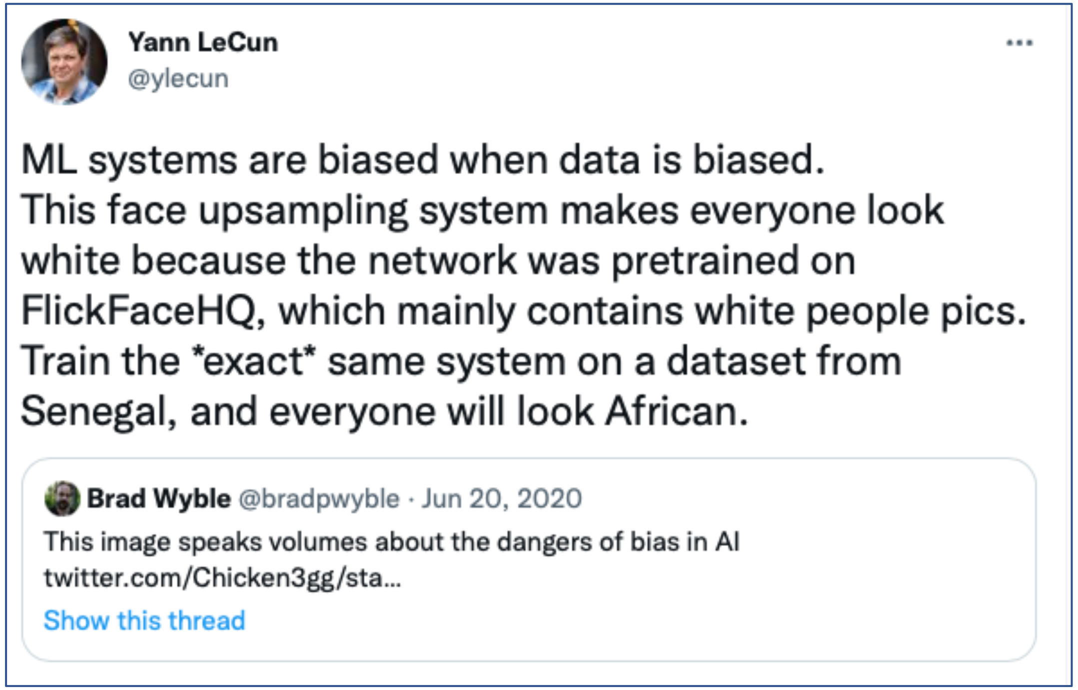
"Data is never this raw, truthful input and never neutral. It is information that has been collected in certain ways by certain actors and institutions for certain reasons."

Catherine D'Ignazio

https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf

Bias in Face Upsampling



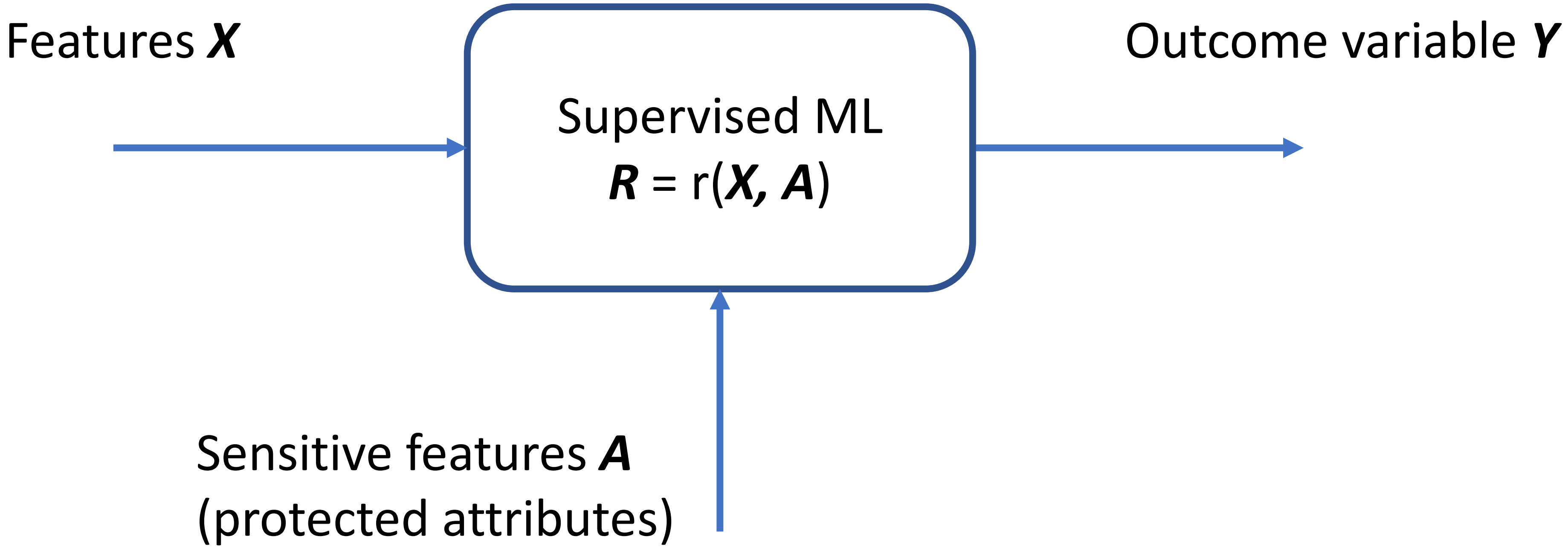


Are AI developer responsible for the creation of unfair AI systems without purposedly designing them to be unfair?

https://twitter.com/ylecun/status/1274782757907030016?lang=en

FAIRNESS CRITERIA

Formal prediction and decision making setting



Fairness through unawareness

Fairness through unawareness if
none of the sensitive features are directly used in the model

Ethnicity	Skills	Experience	Loves tacos	Hired?
Hispanic	Python	1	Yes	No
Hispanic	C++	5	Yes	Yes
White	Java	2	No	Yes
White	C++	3	No	Yes

Supervised ML

Fairness through unawareness

Fairness through unawareness if
none of the sensitive features are directly used in the model

Ethnicity	Skills	Experience	Loves tacos	Hired?
~~Hispanic~~	Python	1	Yes	No
~~Hispanic~~	C++	5	Yes	Yes
~~White~~	Java	2	No	Yes
~~White~~	C++	3	No	Yes

Supervised ML

Fairness through unawareness

Fairness through unawareness if
none of the sensitive features are directly used in the model

Ethnicity	Skills	Experience	Loves tacos	Hired?
~~Hispanic~~	Python	1	Yes	No
~~Hispanic~~	C++	5	Yes	Yes
~~White~~	Java	2	No	Yes
White	C++	3	No	Yes

Inferred

Supervised ML

Sensitive features may still be used indirectly!

Fairness criteria

Independence

$$\boldsymbol{R} \perp \boldsymbol{A}$$

Demographic parity, statistical parity

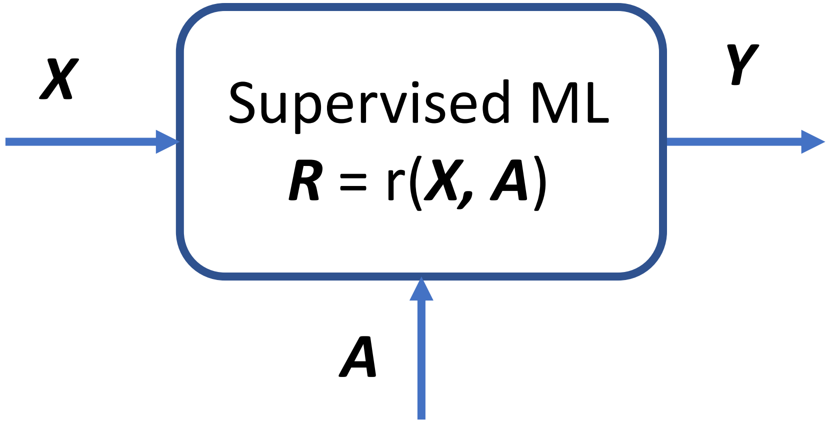
Separation

$$\boldsymbol{R} \perp \boldsymbol{A} \mid \boldsymbol{Y}$$

Equality of odds

Sufficiency

$$\boldsymbol{Y} \perp \boldsymbol{A} \mid \boldsymbol{R}$$



Binary setting

Confusion matrix

		True Class ***Y***	
		positive	negative
Predicted Class ***R***	positive	TP	FP
	negative	FN	TN

Binary setting with sensitive attribute

A=0

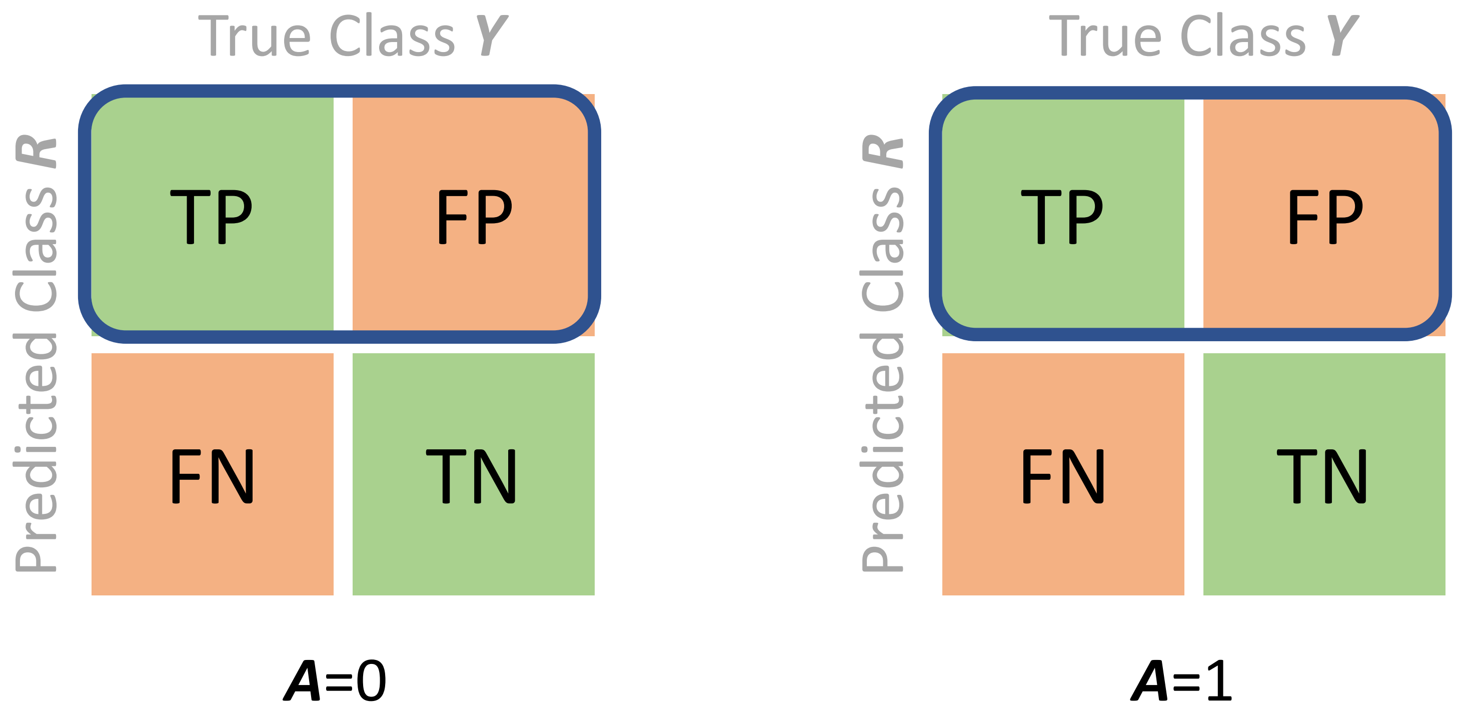
		True Class ***Y***	
		positive	negative
Predicted Class ***R***	positive	TP	FP
	negative	FN	TN

A=1

		True Class ***Y***	
		positive	negative
Predicted Class ***R***	positive	TP	FP
	negative	FN	TN

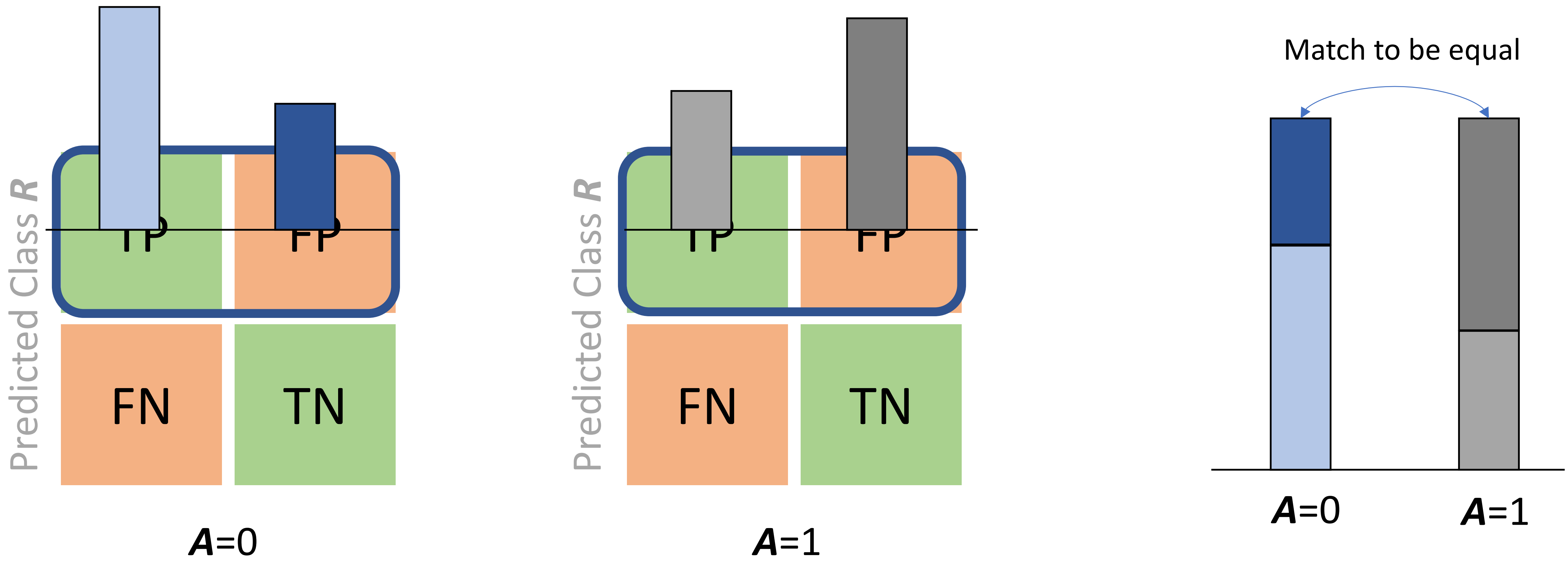
Independence (demographic parity)

$\boldsymbol{R} \perp \boldsymbol{A}$: $P(\boldsymbol{R}=1|\boldsymbol{A}=1) = P(\boldsymbol{R}=1|\boldsymbol{A}=0)$ Positive rate is the same for both groups



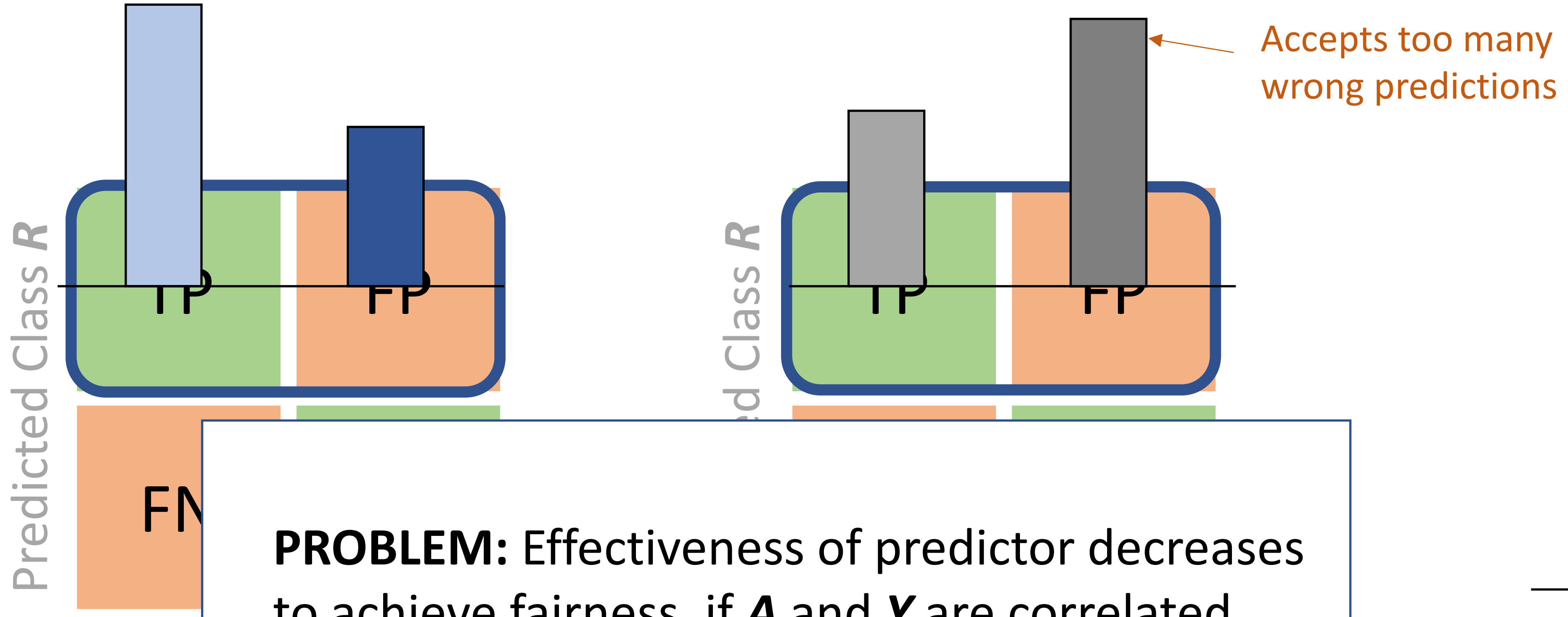
Independence (demographic parity)

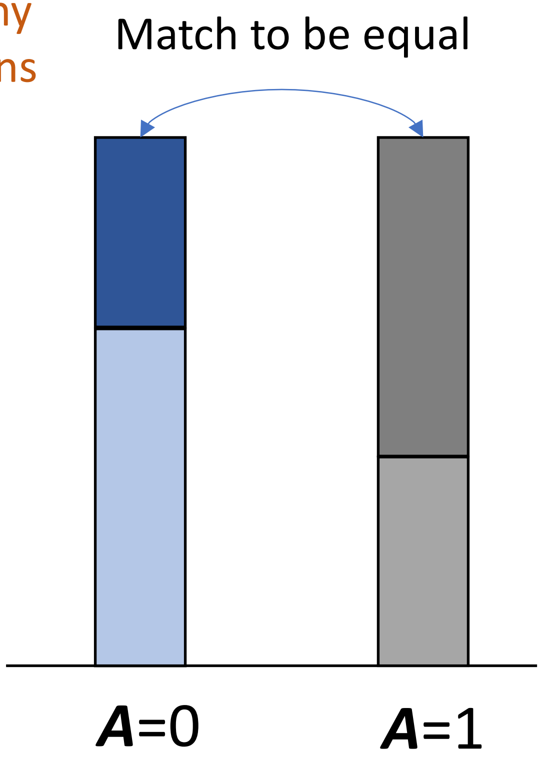
$R \perp A$: $\quad P(R = 1 | A = 1) = P(R = 1 | A = 0)$ Positive rate is the same for both groups



Independence (demographic parity)

$R \perp A$: $P(\boldsymbol{R} = 1|\boldsymbol{A} = 1) = P(\boldsymbol{R} = 1|\boldsymbol{A} = 0)$ Positive rate is the same for both groups





PROBLEM: Effectiveness of predictor decreases to achieve fairness, if ***A*** and ***Y*** are correlated

Separation (equality of odds)

$\boldsymbol{R} \perp \boldsymbol{A} \mid \boldsymbol{Y}$: $\quad P(\boldsymbol{R}=1|\boldsymbol{A}=1,\boldsymbol{Y}) = P(\boldsymbol{R}=1|\boldsymbol{A}=0,\boldsymbol{Y})$

$$\Rightarrow \begin{aligned} P(\boldsymbol{R}=1|\boldsymbol{A}=1,\boldsymbol{Y}=0) &= P(\boldsymbol{R}=1|\boldsymbol{A}=0,\boldsymbol{Y}=0) \\ P(\boldsymbol{R}=1|\boldsymbol{A}=1,\boldsymbol{Y}=1) &= P(\boldsymbol{R}=1|\boldsymbol{A}=0,\boldsymbol{Y}=1) \end{aligned}$$

True Class ***Y***

Predicted Class ***R***		
	TP	FP
	FN	TN

A=0

True Class ***Y***

Predicted Class ***R***		
	TP	FP
	FN	TN

A=1

Separation (equality of odds)

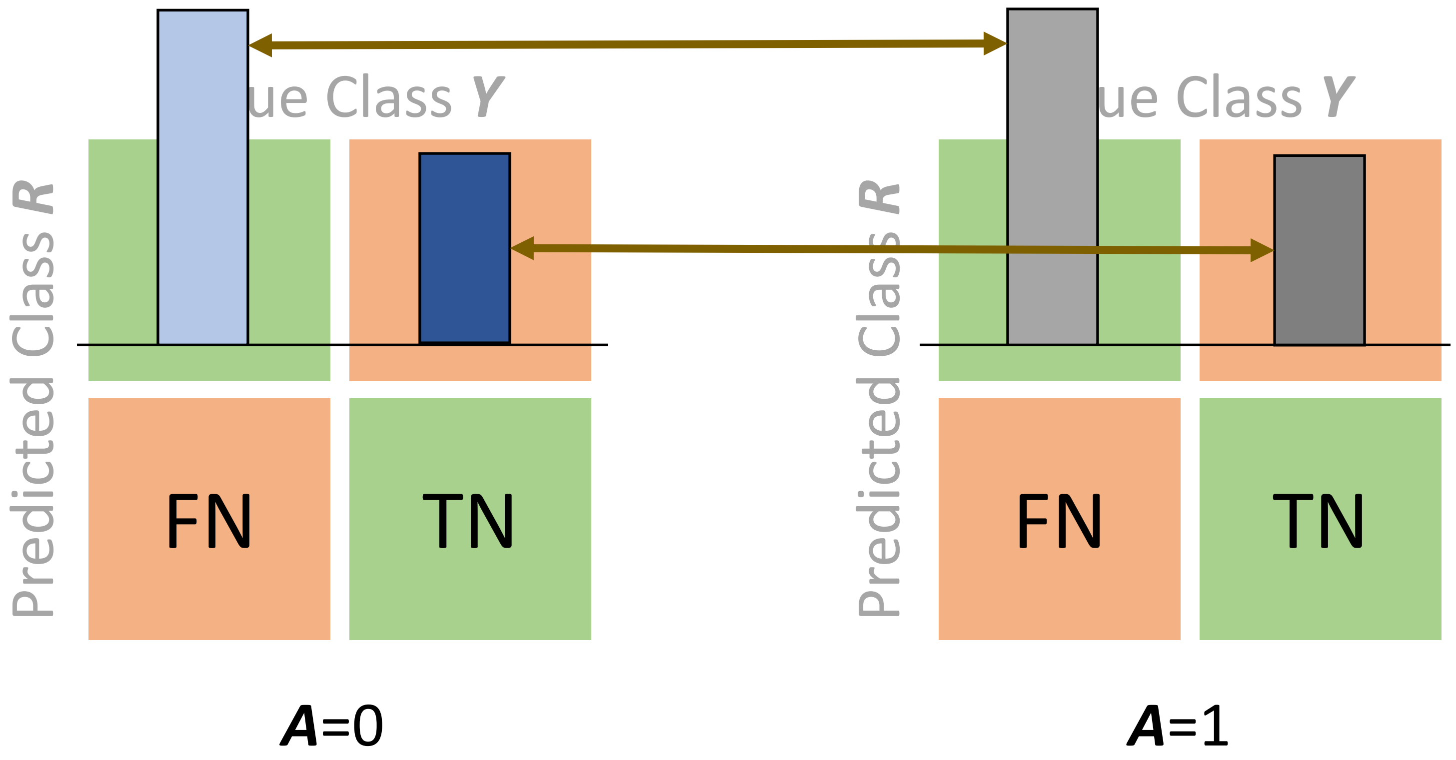
$\boldsymbol{R} \perp \boldsymbol{A} \mid \boldsymbol{Y}$: $\quad P(\boldsymbol{R}=1|\boldsymbol{A}=1,\boldsymbol{Y}) = P(\boldsymbol{R}=1|\boldsymbol{A}=0,\boldsymbol{Y})$

$$P(\boldsymbol{R}=1|\boldsymbol{A}=1,\boldsymbol{Y}=0) = P(\boldsymbol{R}=1|\boldsymbol{A}=0,\boldsymbol{Y}=0)$$

Equal false positive rate

$$P(\boldsymbol{R}=1|\boldsymbol{A}=1,\boldsymbol{Y}=1) = P(\boldsymbol{R}=1|\boldsymbol{A}=0,\boldsymbol{Y}=1)$$

Equal true positive rate

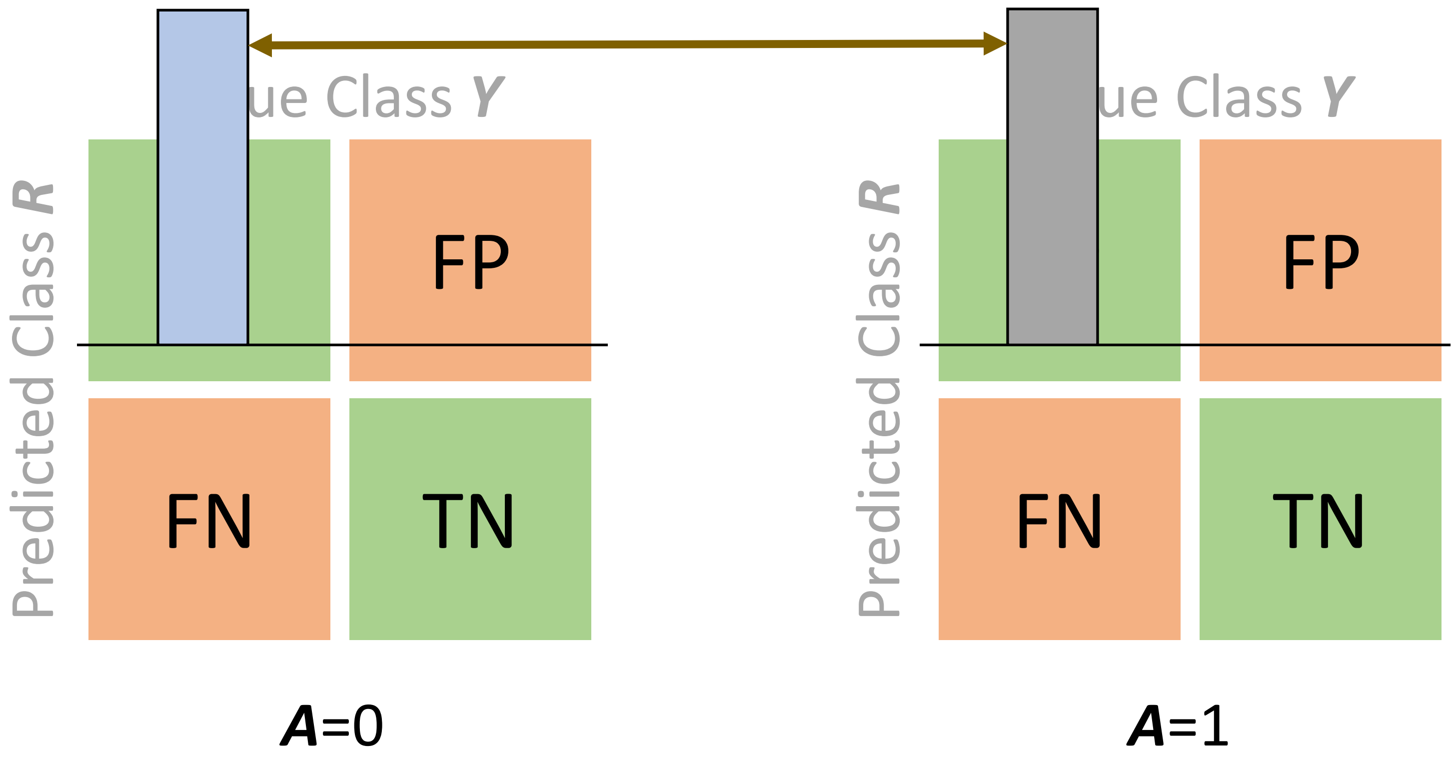


Separation allows correlation between ***R*** and ***A*** to the extent that it is *justified by the target variable **Y***

Equal opportunity

$$P(\boldsymbol{R}=1|\boldsymbol{A}=1,\boldsymbol{Y}=1)=P(\boldsymbol{R}=1|\boldsymbol{A}=0,\boldsymbol{Y}=\mathbf{1})$$

Equal true positive rate



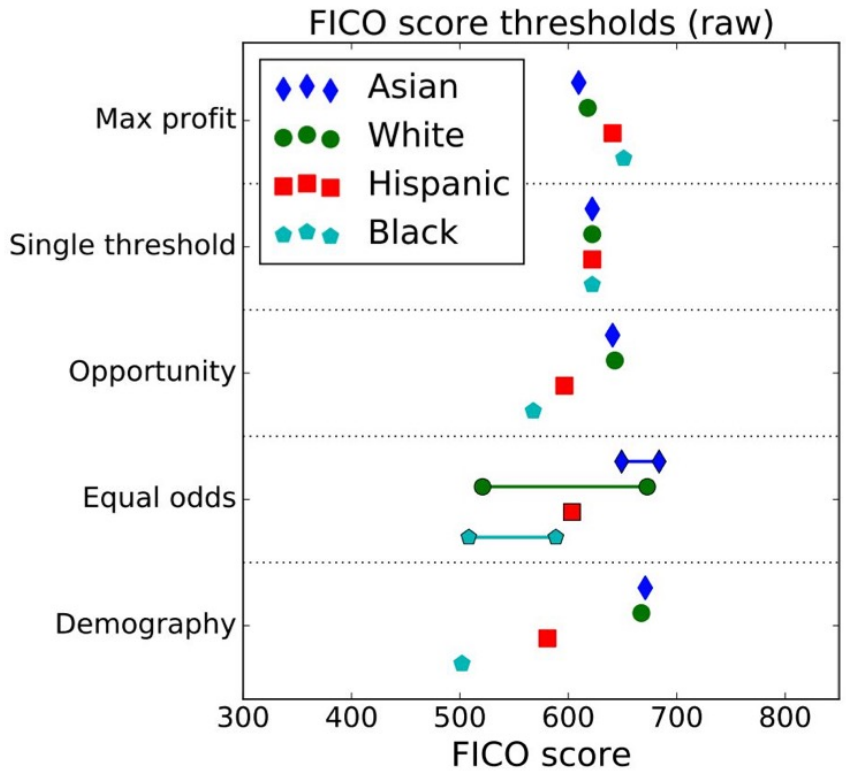
Equal opportunity for people across protected groups.

Bad News

It is not possible to jointly achieve any pair of these conditions

- Independence *xor* Separation
- Independence *xor* Sufficiency
- Separation *xor* Sufficiency

Fairness criteria on their own **cannot** be a "proof of fairness"



Different thresholds for different criteria!

https://arxiv.org/pdf/1610.02413.pdf

Summary of fairness criteria

Fairness	**Criteria**
Unawarness	Exclude $\boldsymbol{A}$ in prediction
Demographic parity	$P(\boldsymbol{R}=1 \mid \boldsymbol{A}=1)=P(\boldsymbol{R}=1 \mid \boldsymbol{A}=0)$
Equality of odds	$P(\boldsymbol{R}=1 \mid \boldsymbol{A}=1, \boldsymbol{Y})=P(\boldsymbol{R}=1 \mid \boldsymbol{A}=0, \boldsymbol{Y})$
Equal opportunity	$P(\boldsymbol{R}=1 \mid \boldsymbol{A}=1, \boldsymbol{Y}=1)=P(\boldsymbol{R}=1 \mid \boldsymbol{A}=0, \boldsymbol{Y}=\mathbf{1})$

METRICS & MITIGATION

Disparity Metrics

Disparity metrics evaluate how far a given predictor departs from satisfying a parity constraint.

Differences & Ratios:

- Demographic Parity Difference / Ratio
- Equal Opportunity Difference / Ratio
- Average Odds Difference / Ratio

Example

Demgraphic Parity:

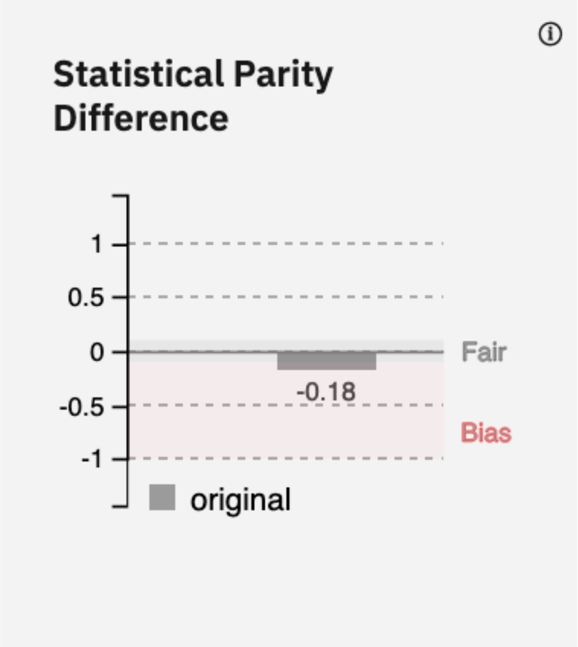
$$P(\boldsymbol{R}=1|\boldsymbol{A}=1)=P(\boldsymbol{R}=1|\boldsymbol{A}=0)$$

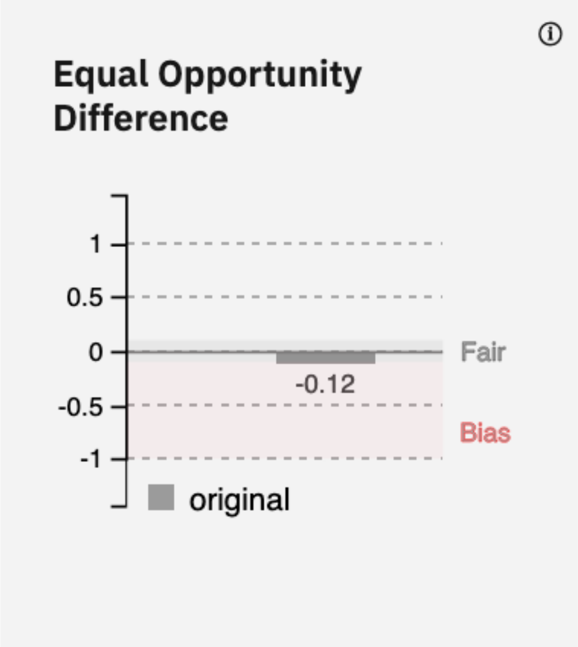
Difference: $\varepsilon = P(\boldsymbol{R}=1|\boldsymbol{A}=1) - P(\boldsymbol{R}=1|\boldsymbol{A}=0)$

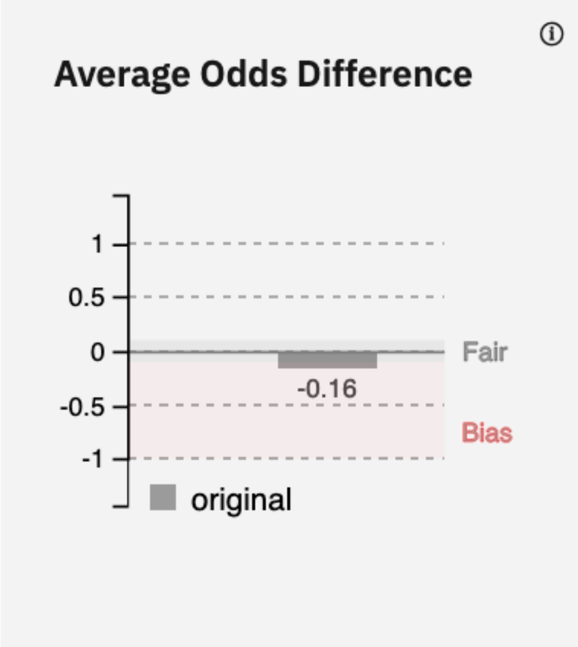
Ratio: $\varepsilon = \dfrac{P(\boldsymbol{R}=1|\boldsymbol{A}=1)}{P(\boldsymbol{R}=1|\boldsymbol{A}=0)} - 1$

Disparity metrics for COMPAS

- Predict a criminal defendant's likelihood of reoffending
- Protected attribute: Race







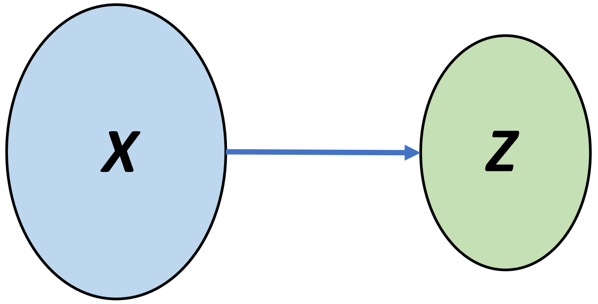
http://aif360.mybluemix.net/data

Mitigation strategies

- Pre-processing
- In-processing
- Post-processing

Mitigation: Pre-processing

1. Adjust the features to be **uncorrelated with the sensitive attribute**
   - “Fair” representation learning
   - Adversarial learning



$Z \perp A$

2. Address **representation bias**
   - Reweighing
   - Oversample from minority groups

3. **Data augmentation**: synthesize data for minority groups
   - Example: from observed “he is a doctor” → synthesize “she is a doctor”

Reweighing

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

$$w = \frac{\text{P expected}}{\text{P observed}}$$

Example: Loan

Expected: P(male) = P(female) = 0.5
P(loan) = P(no loan) = 0.5

P observed	Male	Female
Loan	0.4	0.2
No loan	0.1	0.3

Male loan:

$$w = \frac{0.5 \ * 0.5}{0.4} = 0.625$$

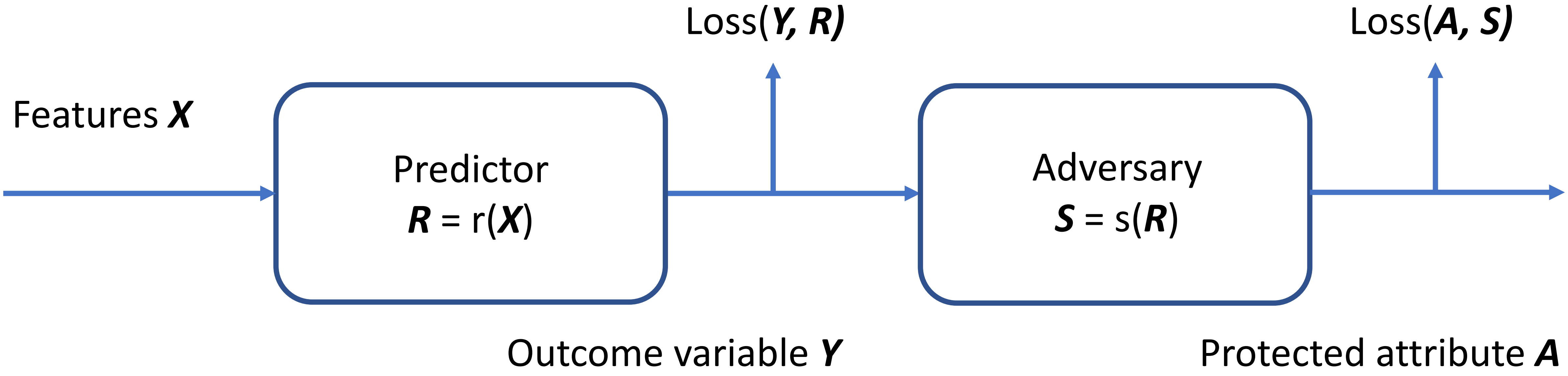
Male no-loan:

$$w = \frac{0.5 \ * 0.5}{0.1} = 2.5$$

https://link.springer.com/content/pdf/10.1007%252Fs10115-011-0463-8.pdf

Mitigation: In-processing

1. Separate model for each protected group ***A***
2. **Adversarial debiasing**
   Maximize prediction accuracy and **simultaneously** reduce an adversary's ability to determine the protected attribute from the predictions



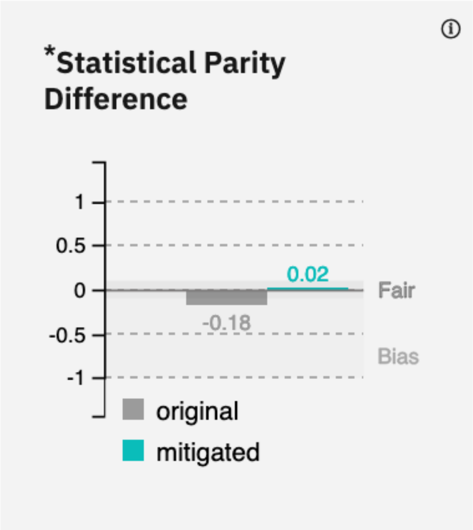
https://arxiv.org/pdf/1801.07593.pdf

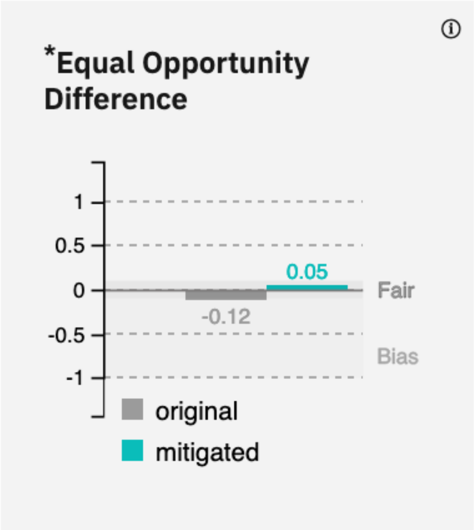
Mitigation for COMPAS

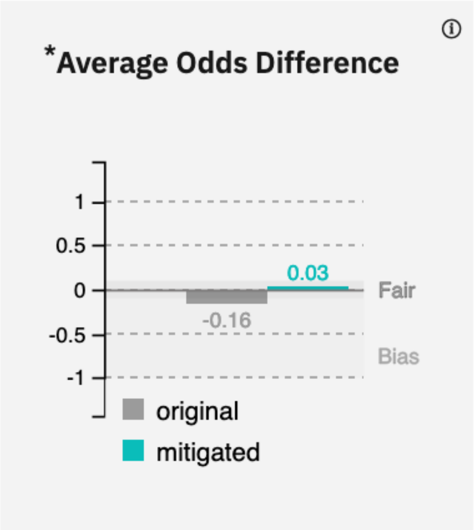
- Pre-processing
  - Reweighting
- In-processing
  - Adversarial debiasing
- Post-Processing
  - Calibrated equalized odds

Mitigation for COMPAS

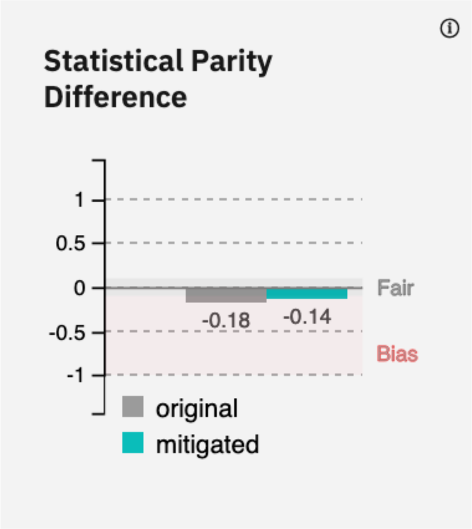
Reweighing

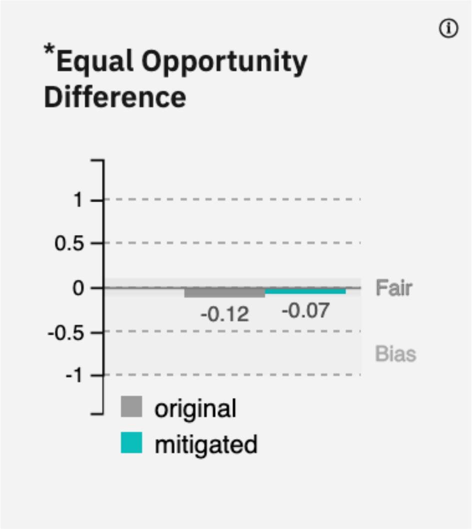


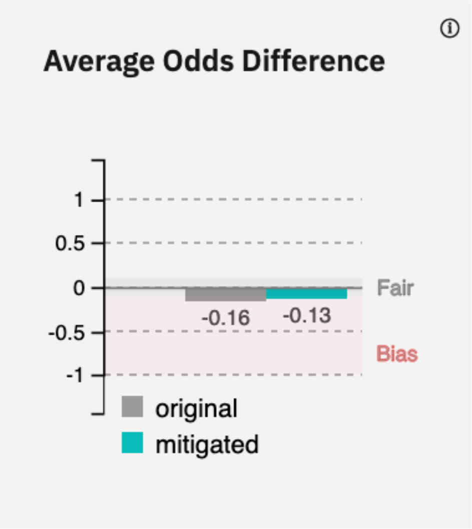




Adversarial debiasing







http://aif360.mybluemix.net/data

Agenda

- Trustworthy AI
- Examples of unfair AI
- Bias & Fairness
- Fairness criteria
- Metrics & Mitigation

Further reading

- Fair ML book (chapter classification)
  https://fairmlbook.org/classification.html
- Tutorial by Moritz Hardt
  https://www.youtube.com/watch?v=Igq_S_7IfOU
- Attacking discrimination with smarter machine learning
  https://research.google.com/bigpicture/attacking-discrimination-in-ml/
- Lecture: Fair, Accountable, and Transparent (FAccT) Deep Learning
  https://hci.stanford.edu/courses/cs335/2020/sp/schedule.html

Thank you for your attention!