Università degli Studi di Padova

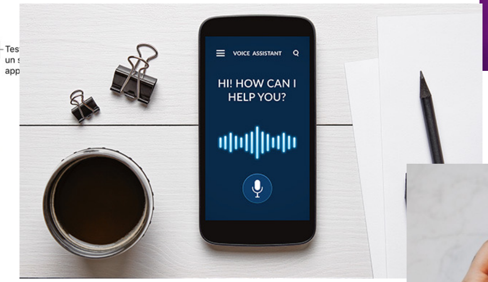
Machine Learning
2024/2025

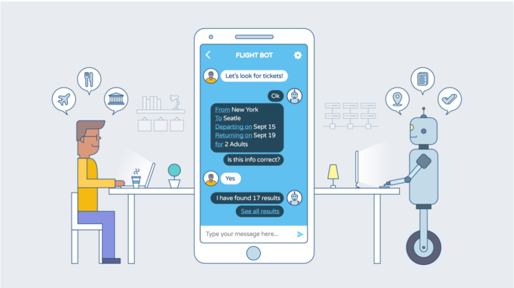
ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CONTROL RESEARCH GROUP

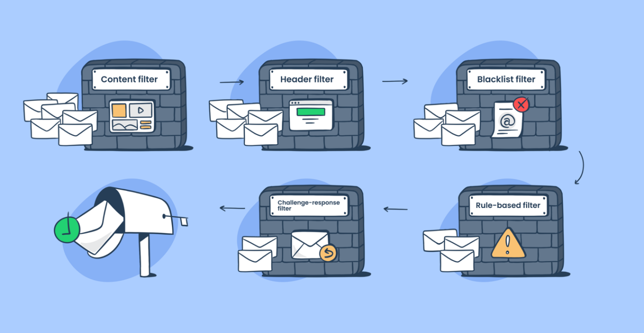
Lecture #29 Algorithmic Fairness

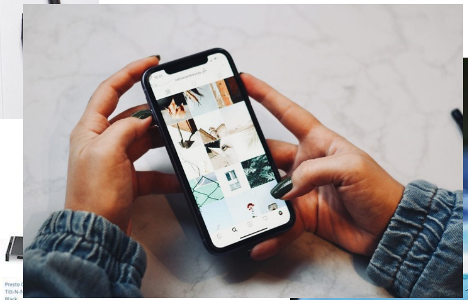
Gian Antonio Susto
Marina Ceccon
* With a special thanks to Alessandro Fabris

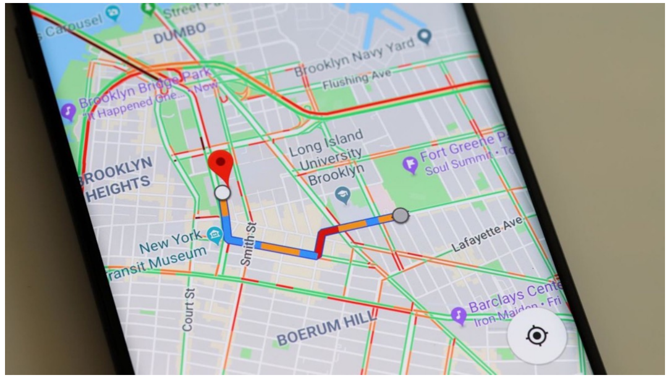
AI is becoming pervasive...

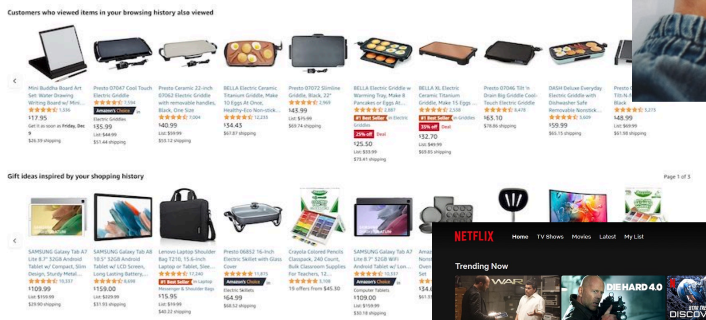


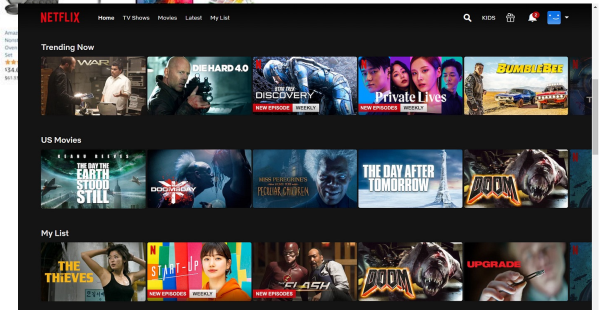


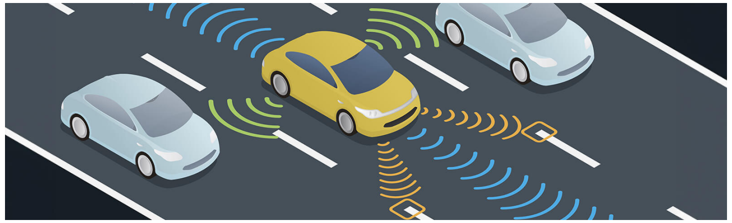












… and accuracy is not enough! Other ML desiderata are required in many applications:

… and accuracy is not enough! Other ML desiderata are required in many applications:

Ideas?

… and accuracy is not enough! Other ML desiderata are required in many applications:

Desiderata of ML systems (beside accuracy):

- Informativeness (for example, confidence intervals or probabilities)
- Robustness (being robust to small data perturbations)
- Scalability (being adaptable to other ‘settings’)
- Sparsity/low complexity
- Fairness
- Explainability/Interpretability

…

… and accuracy is not enough! Other ML desiderata are required in many applications:

Desiderata of ML systems (beside accuracy):

- Informativeness (for example, confidence intervals or probabilities)
- Robustness (being robust to small data perturbations)
- Scalability (being adaptable to other ‘settings’)
- Sparsity/low complexity
- Fairness
- Explainability/Interpretability

…

The more sophisticated the model is, the harder it is to ensure these ‘secondary’ constraints.

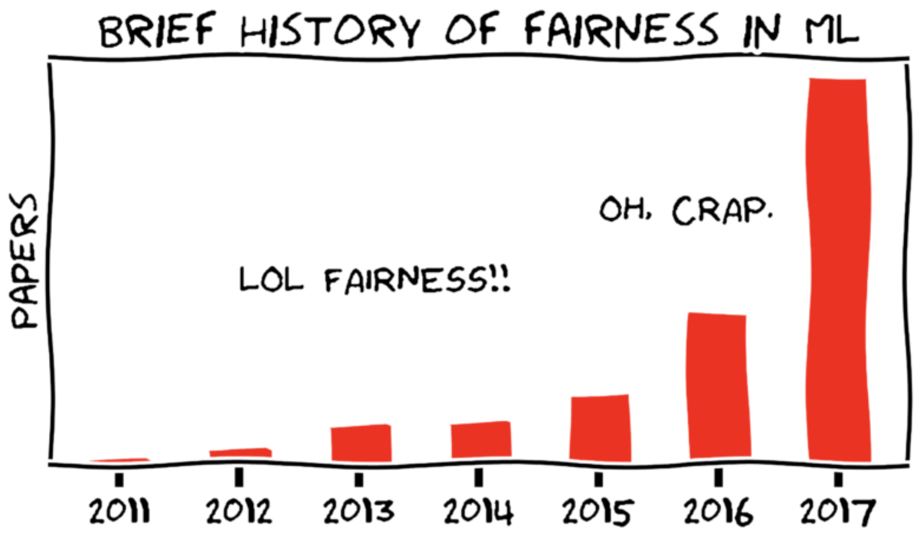
... and accuracy is not enough! Other ML desiderata are required in many applications:

Desiderata of ML systems (beside accuracy):

- Informativeness (for example, confidence intervals or probabilities)
- Robustness (being robust to small data perturbations)
- Scalability (being adaptable to other ‘settings’)
- Sparsity/low complexity
- Fairness (today’s lec.)
- Explainability/Interpretability (lec. 33, 34, 35)

...

The more sophisticated the model is, the harder it is to ensure these ‘secondary’ constraints.



What does it mean to be ‘fair’ in the context of Machine Learning-based technologies?

Ideas?

What does it mean to be ‘fair’ in the context of Machine Learning-based technologies?

When we are dealing with ML-based technologies, we are dealing with ‘automatic’ decision making.

Let’s see first what it means to be <u>unfair</u> in such context with some examples!

Example #01: Fraud Detection

- An automated system to decide who was at risk for fraud detection was recently used in France
- Features that increase the risk score:
  - Low incomes
  - Being unemployed
  - Being a beneficiary of the active solidarity income
  - Living in a “disadvantaged” district
  - Devoting a significant proportion of its income to its rent
  - The lack of work or stable income
  - Being a single parent

https://www.laquadrature.net/2023/11/27/notation-des-allocataires-lindecence-des-pratiques-de-la-caf-desormais-indeniable/

Example #01: Fraud Detection

Do you think these are 'fair' features to use?

- An automated system to decide who was at risk for fraud detection was recently used in France
- Features that increase the risk score:
  - Low incomes
  - Being unemployed
  - Being a beneficiary of the active solidarity income
  - Living in a "disadvantaged" district
  - Devoting a significant proportion of its income to its rent
  - The lack of work or stable income
  - Being a single parent

https://www.laquadrature.net/2023/11/27/notation-des-allocataires-lindecence-des-pratiques-de-la-caf-desormais-indeniable/

Example #02: Job recommendation

Women less likely to be shown ads for high-paid jobs on Google, study shows

theguardian

https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study#:~:text=9%20years%20old-,Women%20less%20likely%20to%20be%20shown%20ads%20for%20high,jobs%20on%20Google%2C%20study%20shows&text=Female%20job%20seekers%20are%20much,than%20men%2C%20researchers%20have%20found.

Example #02: Job recommendation

Why do you this happened?

Women less likely to be shown ads for high-paid jobs on Google, study shows

theguardian

https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study#:~:text=9%20years%20old-,Women%20less%20likely%20to%20be%20shown%20ads%20for%20high,jobs%20on%20Google%2C%20study%20shows&text=Female%20job%20seekers%20are%20much,than%20men%2C%20researchers%20have%20found.

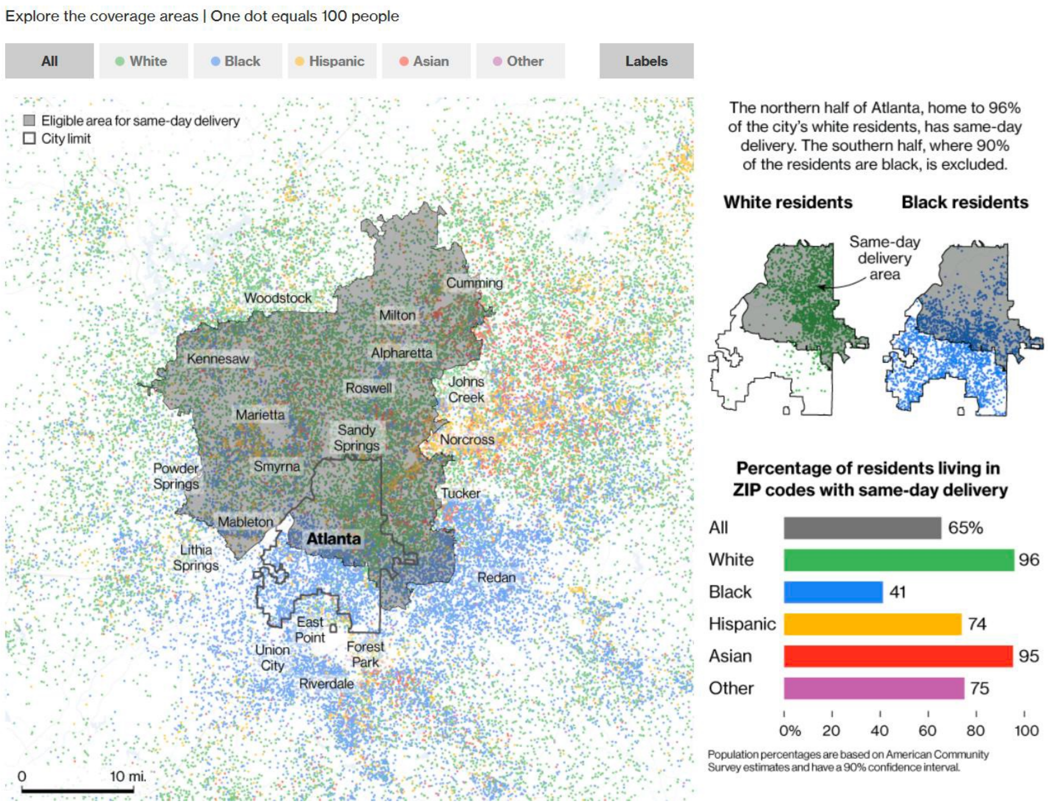
Example #03: Sales & Operations Planning

- In 2016, white U.S. residents were twice as likely to have access to same-day delivery services.
- A machine learning system was deployed to identify zip codes deemed profitable for same-day delivery.

https://www.bloomberg.com/graphics/2016-amazon-same-day/

Example #03: Sales & Operations Planning

- The system prioritized areas near distribution warehouses to optimize logistics.
- Peripheral or remote areas were deprioritized by the algorithm.
- Because zip codes strongly correlate with racial demographics, this resulted in unequal access to services, disproportionately affecting non-white communities.



https://www.bloomberg.com/graphics/2016-amazon-same-day/

Example #04: Smart Mobility

Electric Scooters management – what operators do (Rome example):

- Service Provision: Companies like Lime, Dott, and Bird offer electric scooter-sharing services across Rome.
- Fleet Management: Each operator manages up to 3,000 scooters citywide, as per municipal regulations.
- Operators can move scooters around by logistics teams (vans or small trucks to manually rebalance fleets) or by user incentives: (free rides or credits to users who end trips in low-coverage areas)

Example #04: Smart Mobility

- Moved by profits, all scooters are placed in the city center
- Peripheral areas and people are cut out of the service
- Algorithms to balance profitability and availability can be derived

https://www.ilpost.it/2024/06/12/troppi-monopattini-centro-roma/

st Cerca il POST Shop Regala

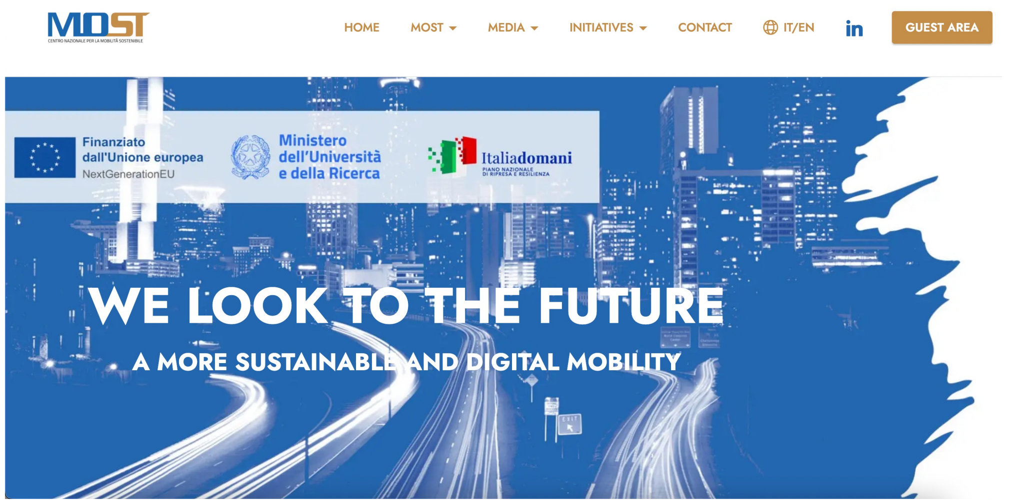
Nel centro di Roma ci sono troppi monopattini

E troppo pochi in periferia, secondo il comune, che continua a multare gli operatori per il mancato rispetto delle regole e ha fermato per una settimana quelli dell'azienda Lime

Condividi

Example #04: Smart Mobility

- Moved by profits, all scooters are placed in the city center
- Peripheral areas and people are cut out of the service
- Algorithms to balance profitability and availability can be derived




Cerca
il POST
Shop Regala

Nel centro di Roma ci sono troppi monopattini

E troppo pochi in periferia, secondo il comune, che continua a multare gli operatori per il mancato rispetto delle regole e ha fermato per una settimana quelli dell'azienda Lime

Condividi

ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CONTROL RESEARCH GROUP

https://www.ilpost.it/2024/06/12/troppi-monopattini-centro-roma/

Cederle, M., Piron, L. V., Ceccon, M., Chiariotti, F., Fabris, A., Fabris, M., & Susto, G. A. (2024). A Fairness-Oriented Reinforcement Learning Approach for the Operation and Control of Shared Micromobility Services. *arXiv preprint arXiv:2403.15780.*

Example #05: RCA Auto

ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CONTROL RESEARCH GROUP

- Algorithms employed by companies in the Italian car insurance industry systematically discriminate against foreign-born drivers.
- In extreme cases, a driver born in Laos may be charged 1,000€ more than a driver born in Milan, all else being equal.

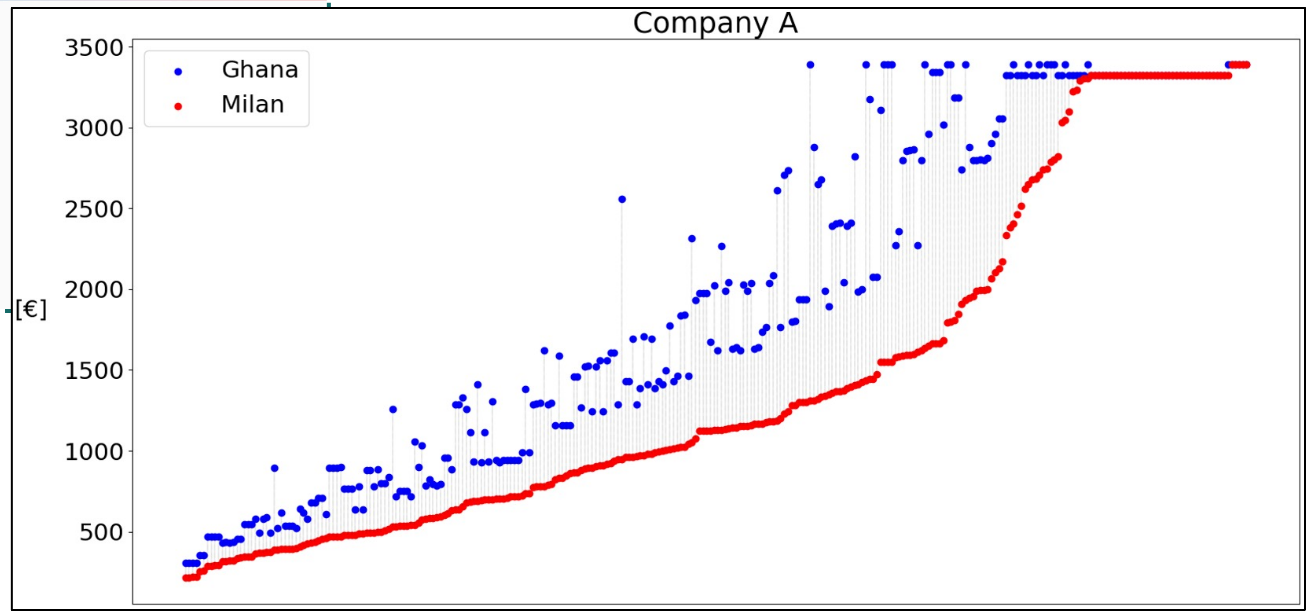
ALGORITHM WATCH

Newsletters Contact Press Deutsch

ABOUT / PROJECTS / PUBLICATIONS / STORIES / POSITIONS / BLOG / DONATE

Costly birthplace: discriminating insurance practice

Two residents in Rome with exactly the same driving history, car, age, profession, and number of years owning a driving license may be charged a different price when purchasing car insurance. Why? Because of their place of birth, according to a recent study.



Fabris et al., Algorithmic Audit of Italian Car Insurance: Evidence of Unfairness in Access and Pricing, https://dl.acm.org/doi/abs/10.1145/3461702.3462569

Example #06: Medical Images (Underrepresentation in datasets)

In medical images datasets, some demographic groups are underrepresented (e.g., females and black people).

Example #06: Medical Images (Underrepresentation in datasets)

Why is that and why it is a problem?

In medical images datasets, some demographic groups are underrepresented (e.g., females and black people).

Example #06: Medical Images (Underrepresentation in datasets)

In medical images datasets, some demographic groups are underrepresented (e.g., females and black people).

Why is that and why it is a problem?

Certain groups may have reduced access to healthcare services, leading to their underrepresentation in medical imaging datasets—this can result in biased diagnostic tools and unequal treatment outcomes

Example #07: Face Recognition (Underrepresentation)

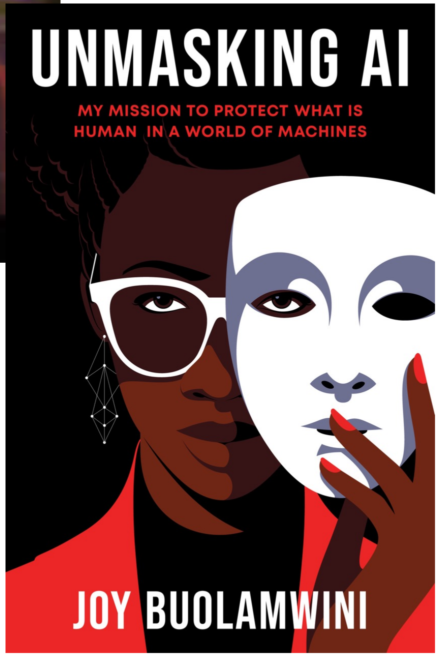
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Example #07: Face Recognition (Underrepresentation)

Why it is a problem?

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%

https://www.youtube.com/watch?v=UG_X_7g63rY



Example #07: Face Recognition (Underrepresentation)

🔍 Case Example: Robert Williams (2020, Michigan, USA)

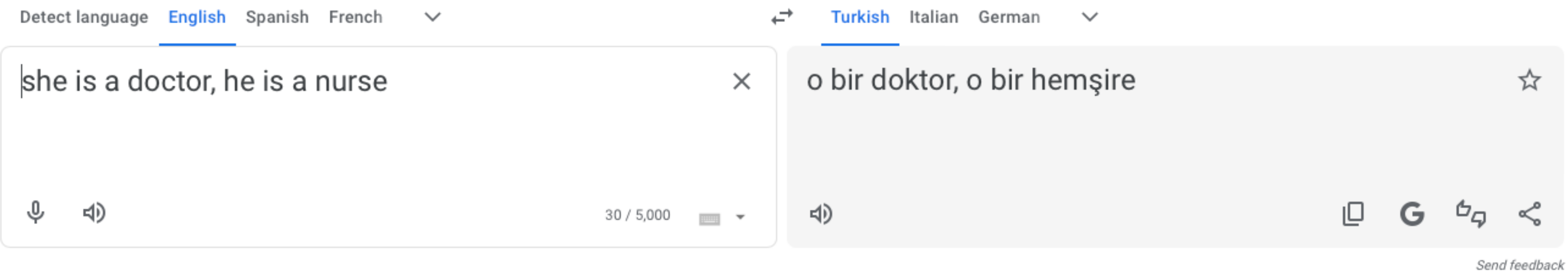
- Robert Williams, a Black man from Detroit, was wrongfully arrested because a facial recognition system falsely matched his face to surveillance footage from a theft.

- The system misidentified him, likely due to racial bias in the training data. His photo was matched despite the video being blurry and low-quality.

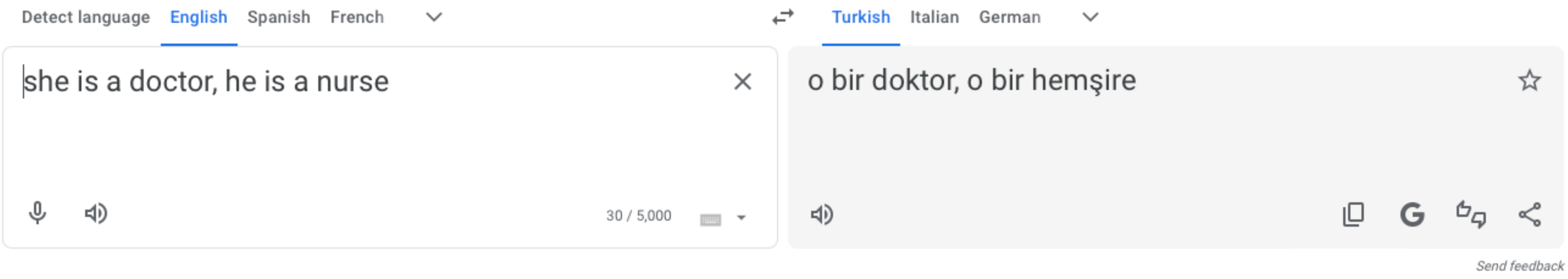
- He was detained for over 30 hours and accused of a crime he didn’t commit. The case was later dropped, and his arrest drew national attention.

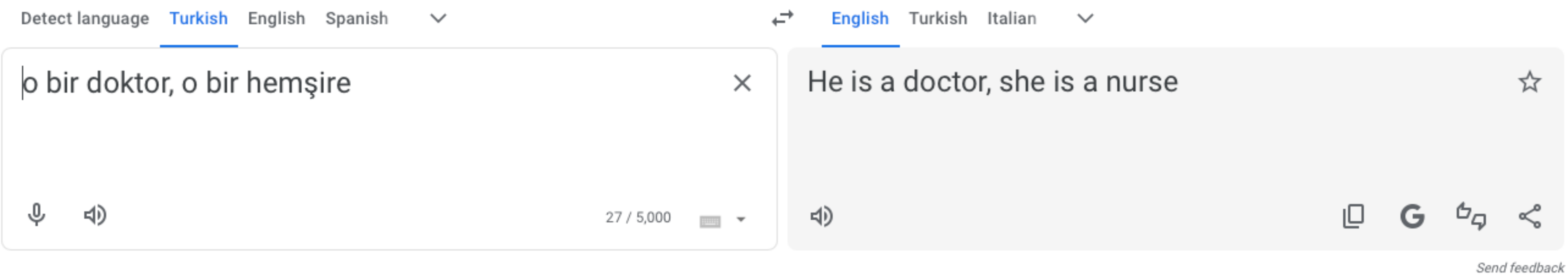
https://quadrangle.michigan.law.umich.edu/issues/winter-2024-2025/flawed-facial-recognition-technology-leads-wrongful-arrest-and-historic

Example #08: Automatic Translators



Example #08: Automatic Translators





Example #09: Generative AI

Prompt: "A doctor is talking to a nurse in a hospital room".

MidJourney

Example #09: Generative AI

Prompt: “A chinese businessperson eats traditional Spanish food in Barcelona”.

MidJourney

Example #09: Generative AI

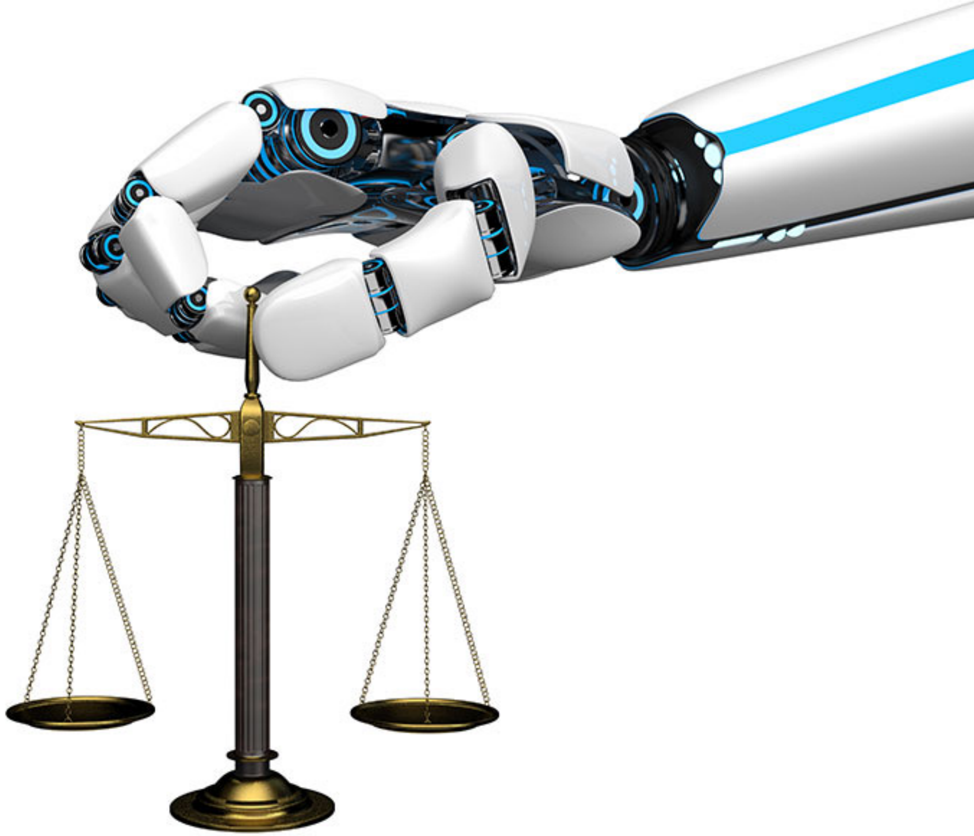
Prompt: “At a hospital in Oslo, a doctor from Ghana talks with a child in the oncology ward”.

MidJourney

Definition of Fairness

- In the context of AI/ML, fairness refers to the absence of bias, discrimination, or favoritism toward individuals or groups in the outcomes, decisions, or processes of AI systems.
- More formally, fairness in AI means that:

An AI system's decisions should not result in unjust or prejudiced outcomes based on sensitive attributes such as race, gender, age, religion, socioeconomic status, or other protected characteristics.



Sensitive/Protected Attributes

- Protected or sensitive attributes in the context of AI are characteristics of individuals that are legally or ethically recognized as grounds on which unfair treatment or discrimination must be avoided.
- Common protected/sensitive attributes include: Race or ethnicity, Gender or sex, Age, Disability, Religion or belief, Sexual orientation, Nationality or immigration status, Marital or family status, Socioeconomic background
- These attributes are considered "sensitive" because they have historically been the basis for discrimination or social inequality. We aim at developing ML-based technologies that do no perpetuate such biases!

Sensitive/Protected Attributes

- Protected or sensitive attributes in the context of AI are characteristics of individuals that are legally or ethically recognized as grounds on which unfair treatment or discrimination must be avoided.
- Common protected/sensitive attributes include: Race or ethnicity, Gender or sex, Age, Disability, Religion or belief, Sexual orientation, Nationality or immigration status, Marital or family status, Socioeconomic background
- These attributes are considered "sensitive" because they have historically been the basis for discrimination or social inequality. We aim at developing ML-based technologies that do no perpetuate such biases!
- Notation:
  - Sensitive attribute: $A \in \mathcal{A}$ (e.g. $\mathcal{A} = \{\text{female, male}\}$ or $\mathcal{A} = \{\text{Black, White, Asian}\}$)
  - Non-sensitive features X, True outcome Y, predicted outcome $\hat{Y}$

Sensitive/Protected Attributes

Why do you think in previous listed examples we have such fairness problems?

- Protected or sensitive attributes in the context of characteristics of individuals that are legally or et grounds on which unfair treatment or discriminat
- Common protected/sensitive attributes include: Gender or sex, Age, Disability, Religion or belief, Sexual orientation, Nationality or immigration status, Marital or family status, Socioeconomic background
- These attributes are considered "sensitive" because they have historically been the basis for discrimination or social inequality. We aim at developing ML-based technologies that do no perpetuate such biases!
- Notation:
  - Sensitive attribute: $A \in \mathcal{A}$ (e.g. $\mathcal{A}$ = {female, male} or $\mathcal{A}$ = {Black, White, Asian})
  - Non-sensitive features X, True outcome Y, predicted outcome $\hat{Y}$

The ‘inertia’ problem in ML fairness

State of the world → Data → Model

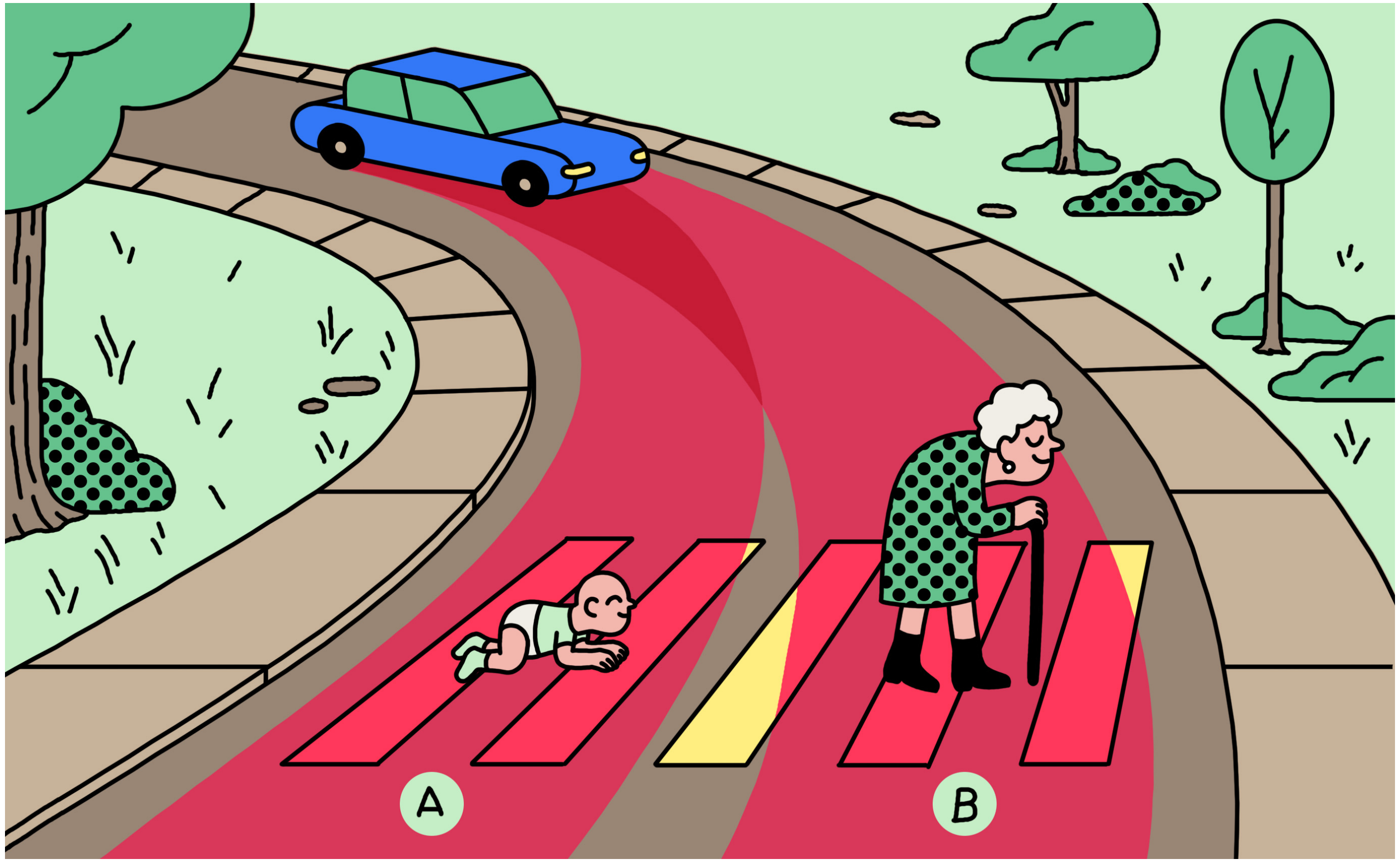
- The discrimination observed in the algorithms is the consequence of the discrimination in the world, and hence in the data.
- Inertia in Machine Learning: there is a delay between the first and the second block.
- While the world is evolving and making progress in reducing discrimination, algorithms are often trained on outdated data, which may not reflect current advancements.

The ‘definition’ problem in ML fairness

- There is no single (or simple) definition of fairness.
  - Some think that the automatic decision should be calibrated on the data, others that it should compensate the discrimination on the data.
- Fairness definitions are somehow related with ethics and will law

The 'definition' problem in ML fairness

- There is no single (or simple) definition of fairness.
  - Some think that the automatic decision should be calibrated on the data, others that it should compensate the discrimination on the data.
- Fairness definitions are somehow related with ethics and will law
- Different cultures/countries follow different ethical principles and therefore have different definitions of what is fair.
- Even within the same culture, what is considered fair may evolve in time.



https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/

Example #10: Recidivism prediction

- COMPAS dataset and algorithm, to predict the probability that a criminal would reoffend.
- An analysis by Pro Publica conducted in 2016 found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.

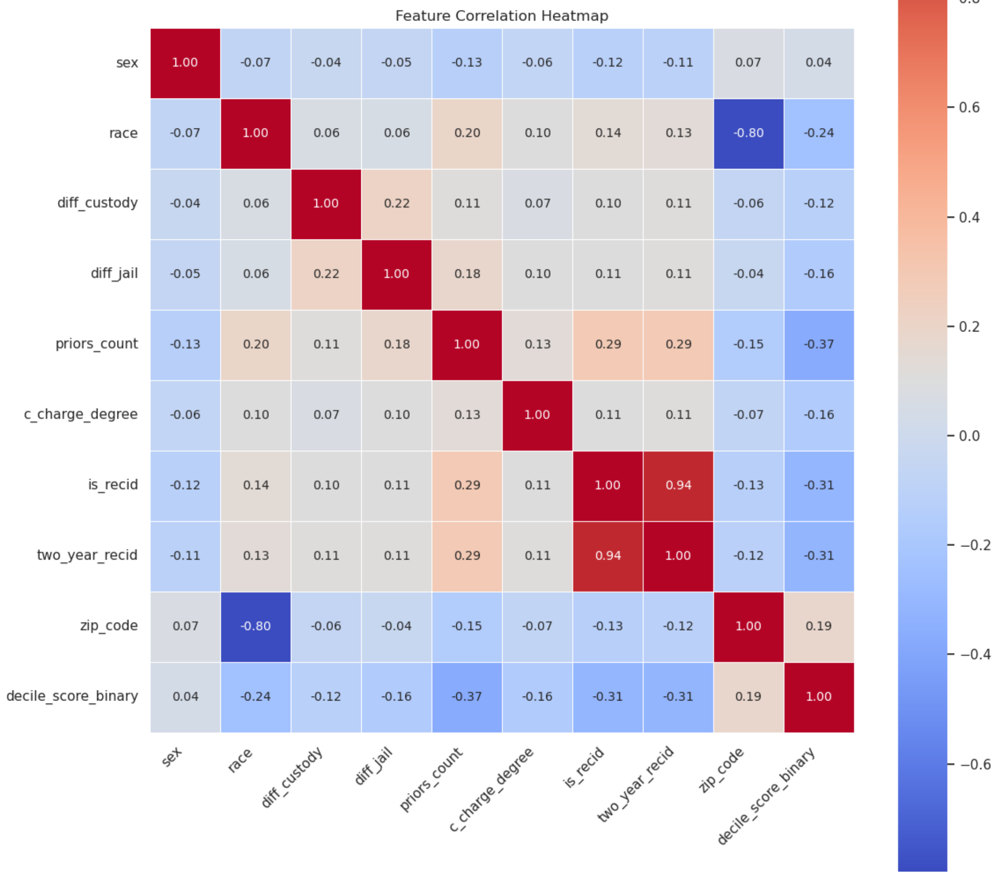
Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

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

Synthetic example

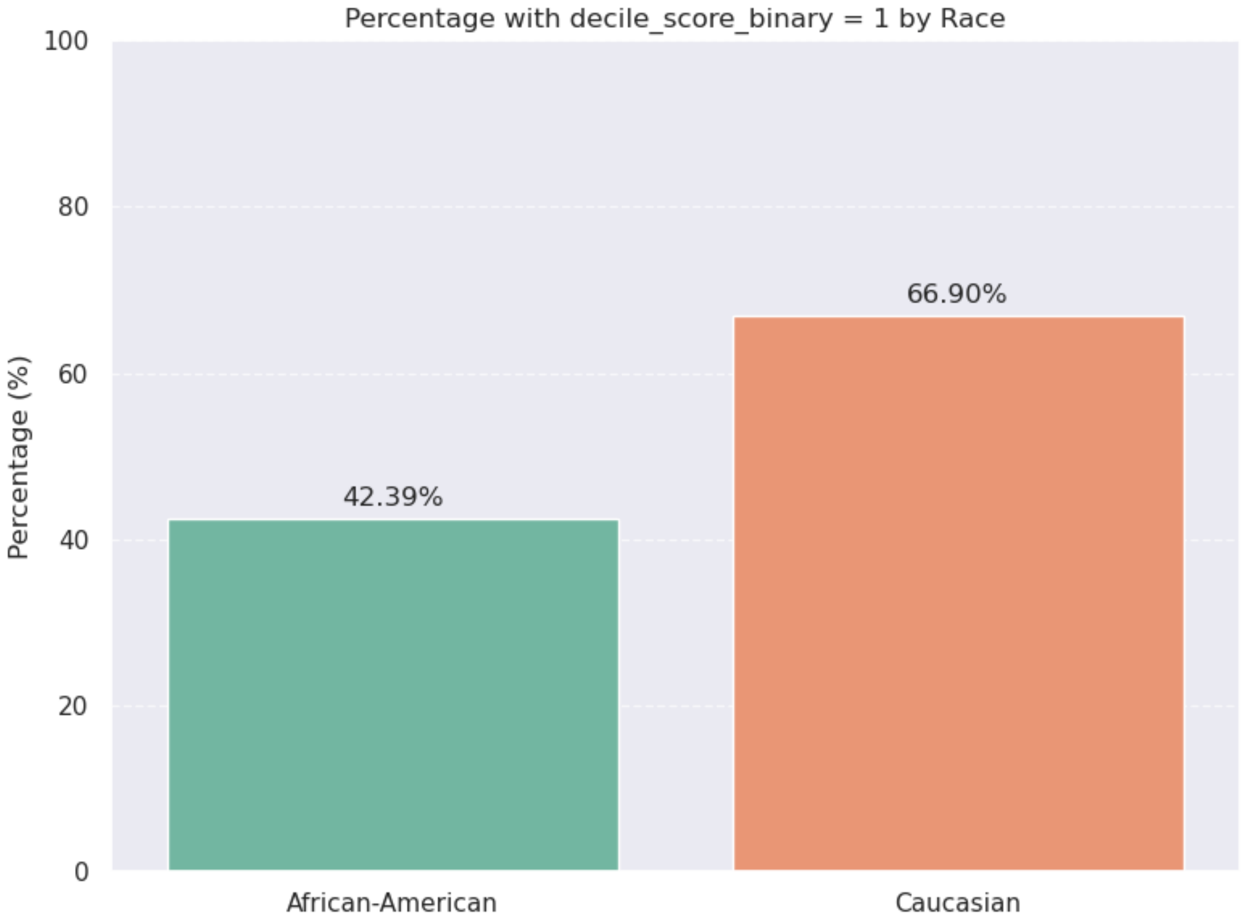
- Let's analyze a synthetic dataset, similar to COMPAS.
- Sex, race and zip code: personal information.
- diff_custody, diff_jail, priors_count, c_charge_degree, is_recid, two_year_recid: how much time they spent in jail, how many crimes they committed, if they reoffended after being arrested the first time…



	sex	race	diff_custody	diff_jail	priors_count	c_charge_degree	is_recid	two_year_recid	zip_code	decile_score_binary
sex	1.00	-0.07	-0.04	-0.05	-0.13	-0.06	-0.12	-0.11	0.07	0.04
race	-0.07	1.00	0.06	0.06	0.20	0.10	0.14	0.13	-0.80	-0.24
diff_custody	-0.04	0.06	1.00	0.22	0.11	0.07	0.10	0.11	-0.06	-0.12
diff_jail	-0.05	0.06	0.22	1.00	0.18	0.10	0.11	0.11	-0.04	-0.16
priors_count	-0.13	0.20	0.11	0.18	1.00	0.13	0.29	0.29	-0.15	-0.37
c_charge_degree	-0.06	0.10	0.07	0.10	0.13	1.00	0.11	0.11	-0.07	-0.16
is_recid	-0.12	0.14	0.10	0.11	0.29	0.11	1.00	0.94	-0.13	-0.31
two_year_recid	-0.11	0.13	0.11	0.11	0.29	0.11	0.94	1.00	-0.12	-0.31
zip_code	0.07	-0.80	-0.06	-0.04	-0.15	-0.07	-0.13	-0.12	1.00	0.19
decile_score_binary	0.04	-0.24	-0.12	-0.16	-0.37	-0.16	-0.31	-0.31	0.19	1.00

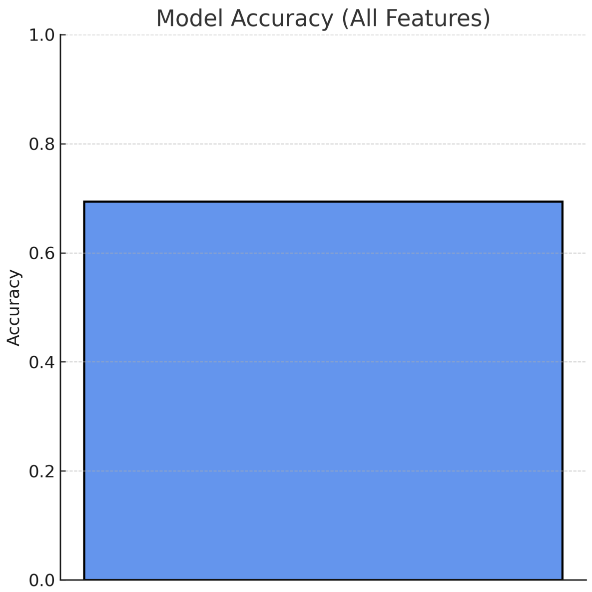
Target distribution

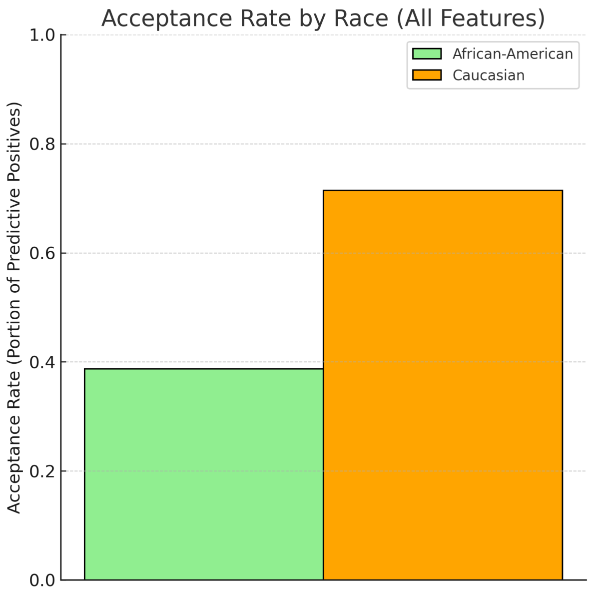
- decile_score_binary: binary target, 1 means low risk of reoffending, 0 means high risk.
- Race = 1 means African-American, Race = 0 means Caucasian.
- The target was determined by human judges.

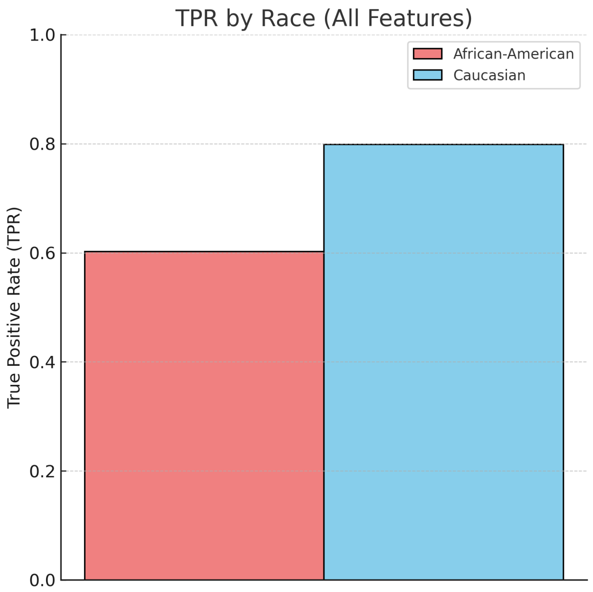


Results on Random Forest Classifier

- We train a Random Forest Classifier to predict decile_score_binary.
- As in COMPAS, African-American are often mispredictd as high risk.







What can we do about it?

Ideas?

What can we do about it?

- Training without considering the variable ‘race’!
- Approach termed as “Fairness through unawarness”.

Themis, greek god of justice

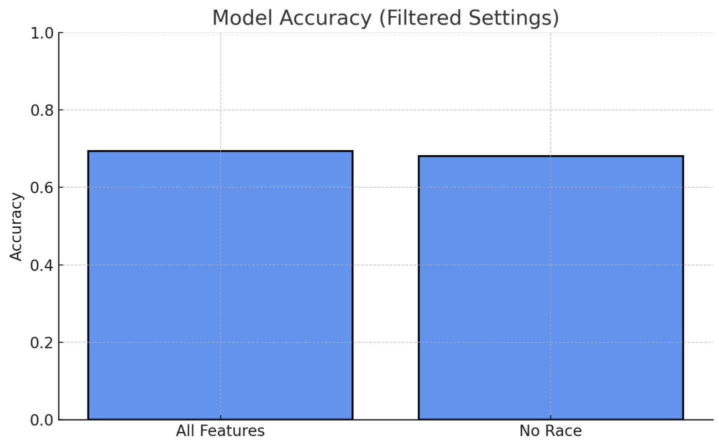
[Fairness metric #01] Fairness through unawareness

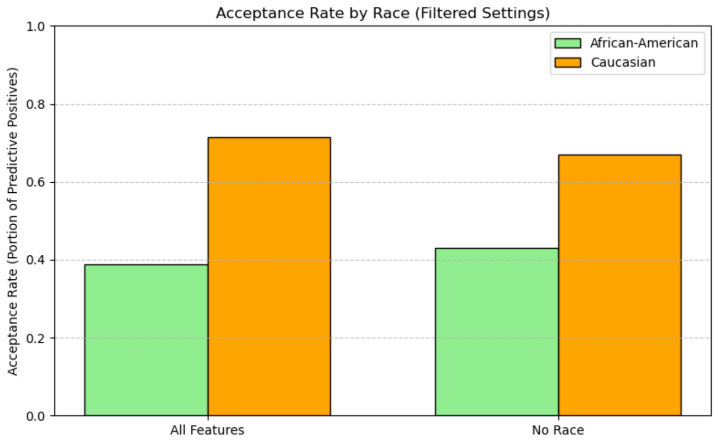
- "An algorithm is fair as long as protected attributes are not explicitly used in the decision-making process."
- Formally, the predictor $\hat{Y} = f(X)$ where $A \notin X$.
- This approach is often appreciated by legal scholars: the *General Data Protection Regulation (GDPR)* prohibits the processing of special categories of personal data, unless specific legal justifications apply.
- The core idea is that an automated system cannot discriminate based on a sensitive attribute if that attribute is not part of the input.

https://certpro.com/gdpr-article-9-sensitive-data-guide/

Training without considering “race”

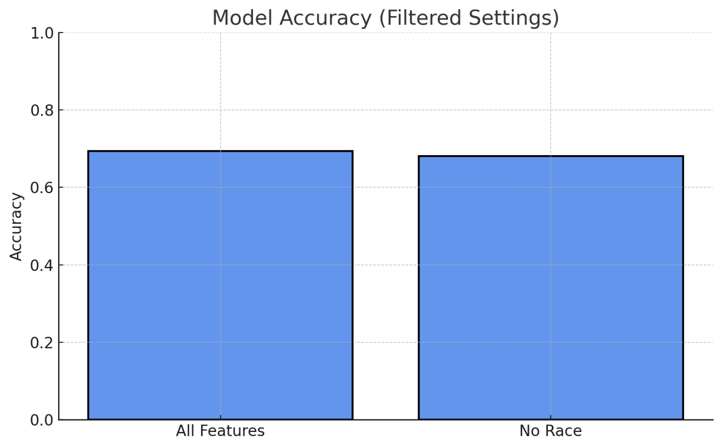
Does it work?

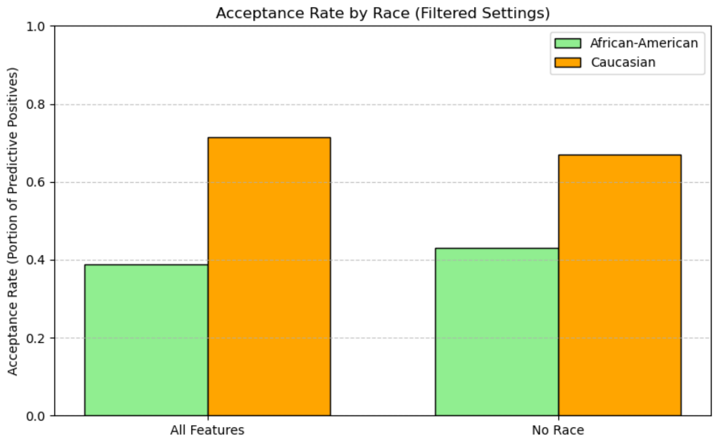


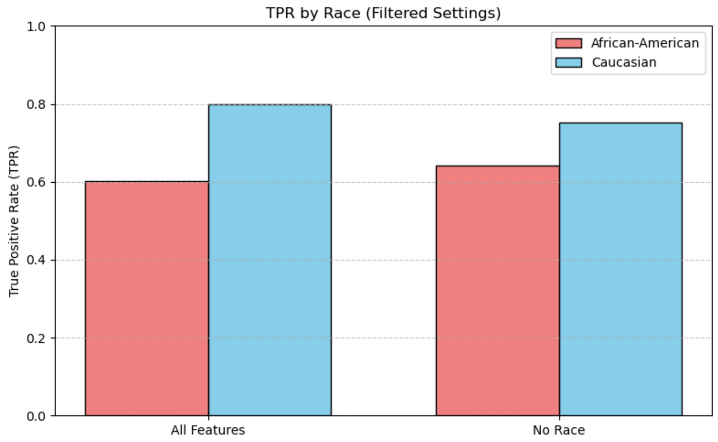


Training without considering “race”

Does it work?







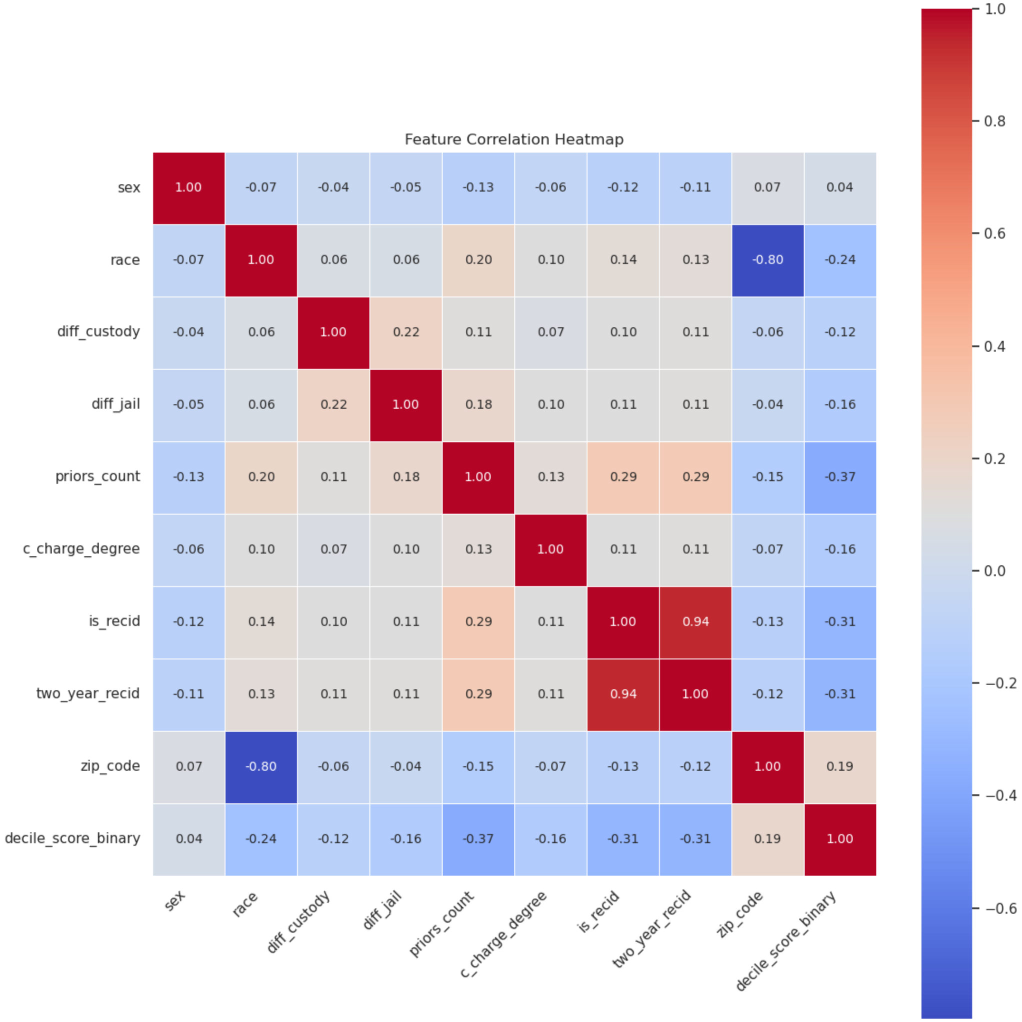
The fairness improves but the gap is still notable.

Can we do better?

- Let’s take a look at the correlation matrix again.

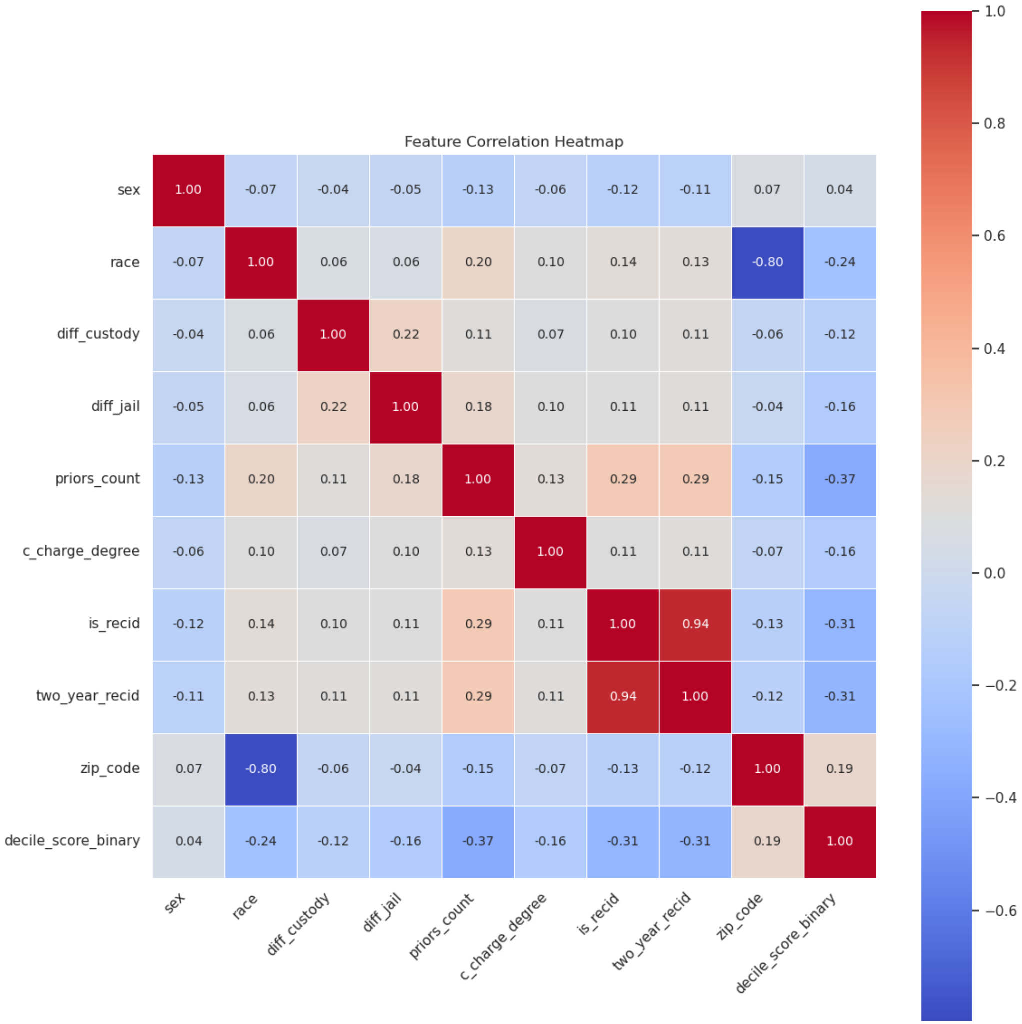
Can we do better?

- Let's take a look at the correlation matrix again.

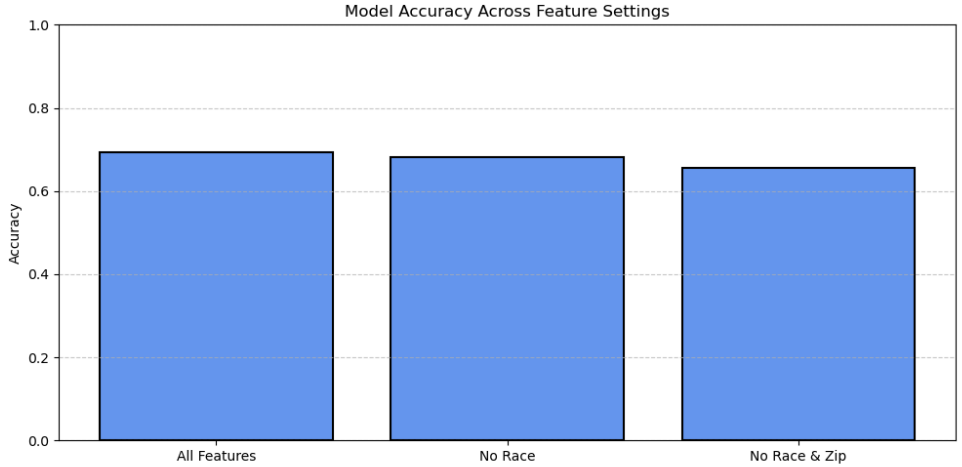


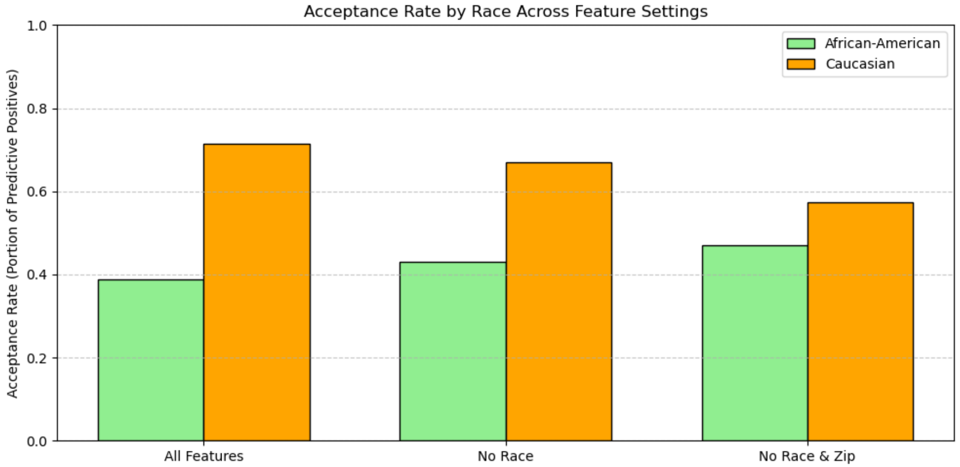
Can we do better?

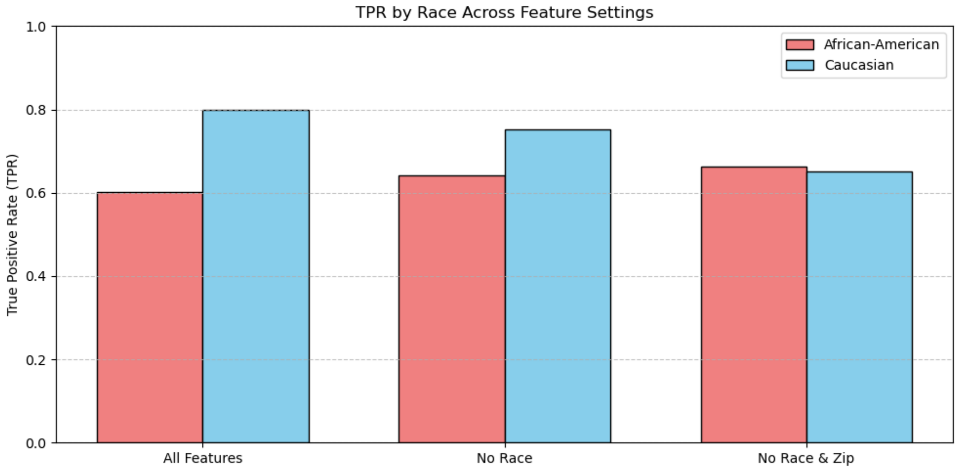
- Let's take a look at the correlation matrix again.
- We observe that **ZIP code** is strongly correlated with **race**.
- In such cases, we say that ZIP code acts as a *proxy* for race.
- Even if race is not explicitly included in the model, it can still be inferred through proxy variables.



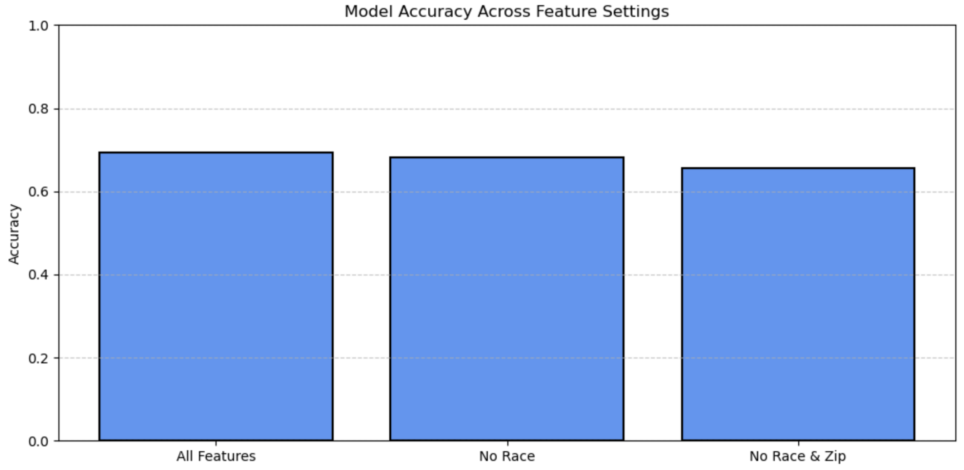
Training without considering “race” and “zip code”

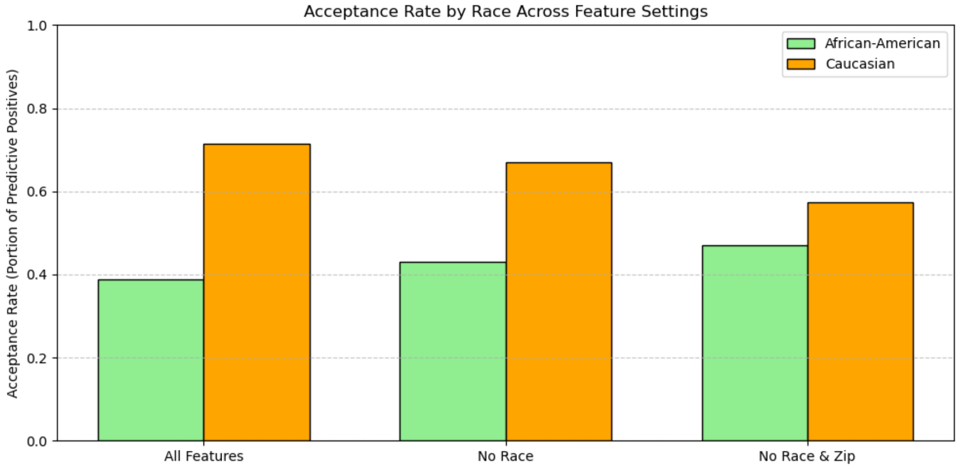


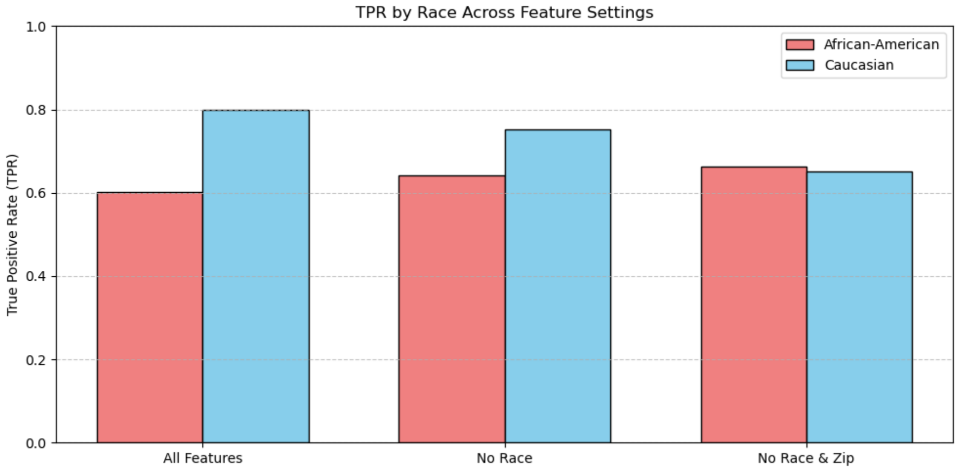




Training without considering “race” and “zip code”







Trade-off between accuracy and fairness!

Fairness metrics

- In the past example we have compared acceptance rates and TPRs.
- These are examples of fairness metrics!

[Fairness metric #02] Demographic Parity

- Used in contexts where a positive outcome is desirable.
- Ideally, we would want the prediction to be independent from the sensitive attribute (gender, ethnicity…)
- In other words, we want the probability of a positive outcome to be as similar as possible across groups.
- Defined as:

$$DP = \mathbb{P}\left(\hat{Y} = 1 \,\middle|\, A = a\right) - \mathbb{P}\left(\hat{Y} = 1 \,\middle|\, A = d\right)$$

- where a is the adavantaged group, while d is the disadvantaged.

[Fairness metric #03] Equality of opportunity

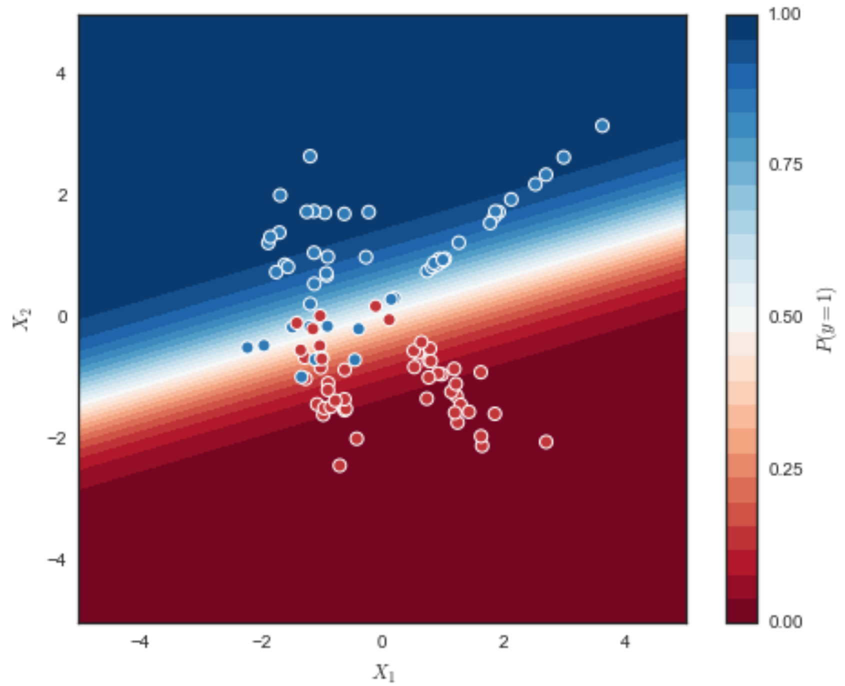
- In other cases, the outcome *should* be dependent on the sensitive attribute.
- Example: in the medical setting, the frequencie of some pathologies depends on gender or ethnicity.
- In this case, we don't aim for similar acceptance rates; instead, we want similar True Positive Rates (TPR).
- Defined as:

$$EO = \mathbb{P}\left(\hat{Y} = 1 \,\middle|\, A = a,\ Y = 1\right) - \mathbb{P}\left(\hat{Y} = 1 \,\middle|\, A = d,\ Y = 1\right)$$

- where a is the adavantaged group, while d is the disadvantaged.

How to 'impose' fairness metrics

- In recidivism prediction, one approach could be to adjust the decision thresholds that determine whether someone is classified as "low risk."
- As discussed in previous lectures, the Random Forest algorithm provides a confidence score for each prediction.
- For the disadvantaged group, lower the threshold for classifying individuals as "low risk" to equalize acceptance rates (or TPRs) across groups.



Other ideas

- Introduce "regularization" constraints in the loss function during training: the model should aim for both accuracy and fairness.
- Pre-process the data to make it more "fair":
  - If the issue is under-representation, consider applying data augmentation techniques.
  - If the problem stems from biased ground truth, look for a more equitable alternative ground truth (e.g., in the past example, "is_recid").
- Alternatively, as we've discussed, remove both the sensitive attribute and any potential proxies for it (*fairness through unawareness)*
- Most of these methods involve a trade-off between accuracy and fairness.

Conclusions

Fairness will be/is required by Law

- The General Data Protection Regulation (GDPR) and the EU AI ACT contain regulations on automated systems, many of which involve fairness constraints.
- It is in the interest of practitioners from several different areas to be able to guarantee it.
- Article 22, Recital 71 of the GDPR: *In order to ensure fair and transparent processing in respect of the data subject, [...] the controller should [...] implement technical and organisational measures appropriate to [...] prevent discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect.*

https://gdpr-info.eu
https://artificialintelligenceact.eu

“Brute-Forcing” Fairness is not always ‘reasonable’

Fairness is, in general, not “solved”

- No single definition of fairness.
- Clear-cut answers about fairness are rare.
- No single software tool will solve fairness across all systems.
- Despite all this, we try to abstract some “general” principles to guide improvement.
- Domain knowledge is fundamental.

However, we are improving...

Prompt: “A doctor is talking to a nurse in a hospital room”.

Adobe Firefly

However, we are improving...

Prompt: “A chinese businessperson eats traditional Spanish food in Barcelona”.

Adobe Firefly

However, we are improving...

Prompt: “At a hospital in Oslo, a doctor from Ghana talks with a child in the oncology ward”.

Adobe Firefly

UNIVERSITÀ DEGLI STUDI DI PADOVA

Machine Learning
2024/2025

AMCO
ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CONTROL RESEARCH GROUP

Thank you!

Gian Antonio Susto