Human-Centered Machine Learning: Measuring Fairness

Dong Nguyen

2024

Utrecht University

recap!

Last time: Intro to fairness

- Dual use
- What do we mean with fairness?
- Harms: Allocative harms, representational harms
- Feedback loops
- Statistical bias and societal bias
- Model development (optimization, evaluation)

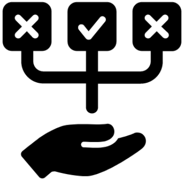
Plan for today

Today: How can we quantify the fairness of ML systems?

- Decision making
- Fairness at the group level

Decision making

Problem setup: decision making



We'll focus on decision making problems framed as *binary* classification tasks:

- Should this person be hired?
- Should this person be admitted to the university?
- Should this person receive parole?

Reminder: Allocative harms.

Human decision making

This is not a new problem!

Eren and Moren found that in the week following an upset loss suffered by the Louisiana State University (LSU) football team, judges imposed sentences that were 7% longer on average. The effect was driven by judges with undergraduate degrees at LSU (emotional impact?).

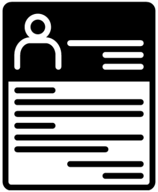
O. Eren and N. Mocan, Emotional Judges and Unlucky Juveniles, American Economic Journal: Applied Economics 10, no. 3 (2018): 171–205. [link]

Human decision making

This is not a new problem!

Example: Fictitious resume with only different names (e.g., gender, white-sounding vs. black-sounding names).

But there are caveats! And in some settings, these tests aren't possible.



See also Chapter 5 ("Testing Discrimination in Practice"); Part 1: Traditional tests for discrimination [link]

For a history of testing, see also 50 Years of Test (Un)fairness: Lessons for Machine Learning, Hutchinson and Mitchell, FAT* 2019 [link]

Anti-discrimination law in the US

Disparate treatment

- *Intentional* discrimination
- Using protected attributes for classification
- Focus on *procedure*

Disparate impact

- *Unintentional* discrimination
- *Unjustified* inequality in outcome
- Focus on *outcome*

Anti-discrimination law in the US

Disparate treatment

- *Intentional* discrimination
- Using protected attributes for classification
- Focus on *procedure*

Disparate impact

- *Unintentional* discrimination
- *Unjustified* inequality in outcome
- Focus on *outcome*

What if knowledge about protected attributes can reduce inequality in outcomes? (remember the example with thresholds)

Protected classes in the US

- race (Civil Rights Act of 1964)
- religion (Civil Rights Act of 1964)
- national origin (Civil Rights Act of 1964)
- sex (Equal Pay Act of 1963 and Civil Rights Act of 1964)
- disability status (Rehabilitation Act of 1973 and Americans with Disabilities Act of 1990)
- ...

Netherlands

Dutch law specifies the following grounds of discrimination:

- race
- sex
- hetero- or homosexual orientation
- political opinion
- religion
- belief
- disability or chronic illness
- civil status
- age
- nationality
- working hours (full time or part time)
- type of contract (temporary or permanent)

Source: https://www.government.nl/topics/discrimination/prohibition-of-discrimination

~~Fairness through unawareness?~~

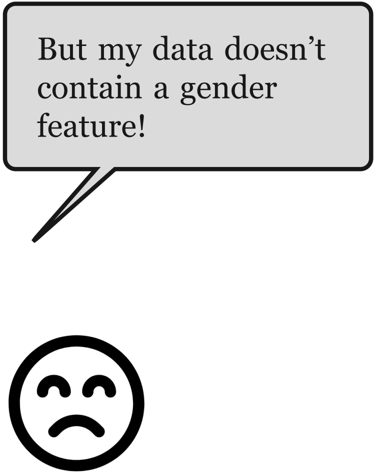
But my data doesn’t contain a gender feature!

~~Fairness through unawareness?~~

But my data doesn't contain a gender feature!

Why is leaving out sensitive features not a solution?

~~Fairness through unawareness?~~



The remaining features may *correlate* with the sensitive features. This is often the case with large features spaces (most of modern ML!)

E.g., proxies (zip code for race)

~~Fairness through unawareness?~~

But my data doesn't contain a gender feature!

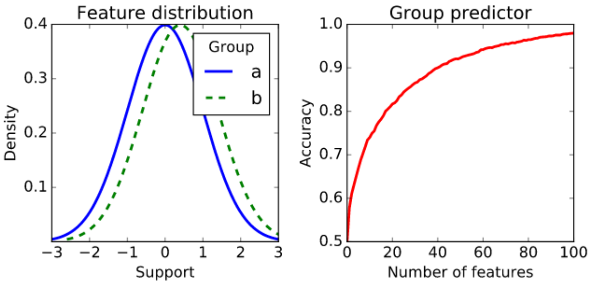


Figure: Fig. 4 from FairML book, "Classification"

~~Fairness through unawareness?~~

But my data doesn't contain a gender feature!

Amazon ditched AI recruiting tool that favored men for technical jobs

"[..] It penalized résumés that included the word "women's", as in "women's chess club captain". And it downgraded graduates of two all-women's colleges, according to people familiar with the matter."

https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

(11 Oct 2018)

Problem setup

- Features: X
- Target variable/outcome: Y, e.g. {0,1} with binary classification
- We want to predict Y from X
- Often we have a score function R = r(X)
- We make a decision based on a threshold: $D = \mathbb{1}\{R > t\}$
- We have a sensitive attribute $A \in \{a, b\}$ (assuming two groups).

Problem setup

- Features: X
- Target variable/outcome: Y, e.g. {0,1} with binary classification
- We want to predict Y from X
- Often we have a score function R = r(X)
- We make a decision based on a threshold: $D = \mathbb{1}\{R > t\}$
- We have a sensitive attribute $A \in \{a, b\}$ (assuming two groups).

Note: We'll use '*decision*' and '*prediction*' interchangeably.

Problem setup

- Features: X
- Target variable/outcome: Y, e.g. {0,1} with binary classification
- We want to predict Y from X
- Often we have a score function R = r(X)
- We make a decision based on a threshold: $D = \mathbb{1}\{R > t\}$
- We have a sensitive attribute $A \in \{a, b\}$ (assuming two groups).

Should I give this person a loan?

- Features: income, debt, …
- Y: Will this person repay their loan? (1=yes, 0=no)
- D: Provide loan (1=yes, 0=no)
- A ∈ {male, female}

Confusion matrix

		Outcome (Y)	
		$(+)$	$(-)$
Decision (D)	$(+)$	$\mathrm{TP} = 5$	$\mathrm{FP} = 2$
	$(-)$	$\mathrm{FN} = 3$	$\mathrm{TN} = 5$

TP = true positive;
FP = false positive;
FN = false negative;
TN = true negative

True positive rate / Recall:
$P[D = +|Y = +] = \frac{TP}{TP+FN}$

False positive rate:
$P[D = +|Y = -] = \frac{FP}{FP+TN}$

True negative rate:
$P[D = -|Y = -] = \frac{TN}{FP+TN}$

False negative rate:
$P[D = -|Y = +] = \frac{FN}{TP+FN}$

Confusion matrix

		Pays back loan (Y)	
		(+)	(−)
Provide loan (D)	(+)	$\mathrm{TP} = 5$	$\mathrm{FP} = 2$
	(−)	$\mathrm{FN} = 3$	$\mathrm{TN} = 5$

TP = true positive;
FP = false positive;
FN = false negative;
TN = true negative

Different stakeholders have different goals.

What would applicants find important? And what about the bank?

Type of errors

Suppose we're building a system to judge whether someone is guilty (guilty=1; innocent=0). Which type of error is more problematic? *False negatives* or *false positives*.

Type of errors

Suppose we're building a system to judge whether someone is guilty (guilty=1; innocent=0). Which type of error is more problematic? *False negatives* or *false positives*.

False positive: Innocent person is judged guilty (and perhaps send to prison)
False negative: Guilty person is judged innocent (and released).

Plan for today

There is not one best way of measuring "fairness".

Terminology: privileged group, majority group (doesn't need to be the same, but often is).

Today: How can we quantify the fairness of ML systems?

- Decision making
- Fairness at the group level

Measuring fairness: Groups

Measuring fairness at the level of groups

Do outcomes systematically differ between different groups?

Three criteria:

equal decision measures	conditional on outcome	conditional on decision
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

Measuring fairness at the level of groups

Do outcomes systematically differ between different groups?

Three criteria:

equal decision measures	conditional on outcome	conditional on decision
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

Equal decision measures

$A \in \{a, b\}$ sensitive attribute; D is the decision

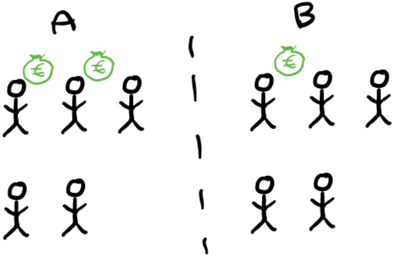
$$A \perp D$$

A generalization is: $A \perp R$.
In a binary classification scenario (e.g., $D = 1$ means hire this person):

$$P[D = 1 | A = a] = P[D = 1 | A = b]$$

The actual outcome is *not considered*
Also called: *demographic parity* or *statistical parity*.

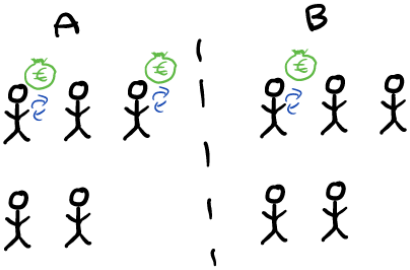
Equal decision measures



If group ***A*** and group ***B*** both apply for a loan at your bank, this is satisfied if an equal % applicants of group ***A*** and % applicants of group ***B*** are granted a loan. (Regardless of whether one group is more likely to repay.)

Here: *no*,
because: A: 2/5=0.4 vs. B: 1/5=0.2

Equal decision measures



Now, what if this classifier makes “no errors”, $(D = Y)$?

That is, all applicants who are selected indeed repay their loan and all others indeed would not have repaid their loan.

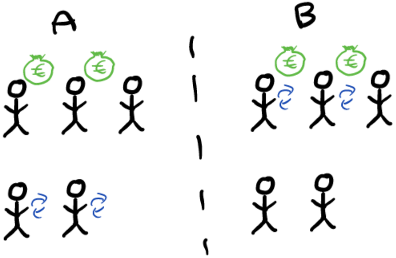
Statistical parity would not be satisfied!

Equal decision measures

Ignores the true outcome Y. Doesn't take "merit" of individuals into account. Why would we want this?

- It might very difficult or impossible to measure the actual outcome.
- We may believe that the observed relation between the attributes and outcome is unfair (e.g. historical prejudice).

Equal decision measures



Caveat: Statistical parity can be satisfied while procedure is unfair.

E.g. having high accuracy in one group, and random predictions in the other group (as long as decision rates are equal).

Equal decision measures

We can relax this with a slack parameter:

$$|P[D = 1|A = a] - P[D = 1|A = b]| <= \epsilon$$

Or we could look at the ratio (a =unprivileged / b=privileged):

$$\frac{P[D = 1|A = a]}{P[D = 1|A = b]}$$

Relates to 80 percent rule in disparate impact law.
Example: Of the men applying at your company, you accept 60%. Of the women applying, you accept 30%. So: $0.3/0.6 = 0.5$, which is < 0.8.

Equal decision measures: Conditional statistical parity

One relaxation is **conditional statistical parity** by controlling for a set of *legitimate* attributes. For example, acceptance rate should be equal across different groups when *controlling* for education.

$A \in \{a, b\}$ sensitive attribute; D is the decision; E is the legitimate sensitive attribute.

In a binary classification scenario (e.g., $D = 1$ means hire this person, $E = e$ means university education):

$$P[D = 1|E = e, A = a] = P[D = 1|E = e, A = b]$$

Algorithmic decision making and the cost of fairness, Corbett-Davies et al., KDD '17 [link]

Equal decision measures: Conditional statistical parity

feature	group	outcome	prediction
E1	F	0	1
E1	F	1	0
E2	F	1	1
E1	M	1	1
E1	M	0	0
E2	M	1	1
E2	M	1	1

Statistical parity? Conditional statistical parity?

Equal decision measures: Conditional statistical parity

feature	group	outcome	prediction
E1	F	0	1
E1	F	1	0
E2	F	1	1
E1	M	1	1
E1	M	0	0
E2	M	1	1
E2	M	1	1

Statistical parity? Conditional statistical parity? Statistical parity: F= 2/3; M=3/4. → no! Conditional statistical parity: When E=E1: F=0.5; M=0.5; When E=E2: F=1, M=1. → yes!

Equal decision measures: Conditional statistical parity

Key open question:

Which attributes are legitimate sources of discrimination?

Measuring fairness at the level of groups

Do outcomes systematically differ between different groups?

Three criteria:

equal decision measures	**conditional on outcome**	**conditional on decision**
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on outcome

Informally: People with the same outcome should be treated the same.

$A \in \{a, b\}$ sensitive attribute; D is the decision; Y is the outcome

$$D \perp A | Y$$

A generalization is: $R \perp A | Y$.

In a binary classification setting: $D \perp A | Y = 1$ and $D \perp A | Y = 0$

Conditional on outcome

True positive rates/recall **(equal opportunity)**:

$$P[D = 1|Y = 1, A = a] = P[D = 1|Y = 1, A = b]$$

Example: Everyone who will repay a loan should have the same likelihood of receiving a loan (regardless of the sensitive attribute).

False positive rates:

$$P[D = 1|Y = 0, A = a] = P[D = 1|Y = 0, A = b]$$

Both constraints: **equalized odds**

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on outcome

We need to know the (true) outcomes!
Often, it's hard or impossible to know the true outcomes.

- Hiring
- University admission
- ...

Conditional on outcome

True positive rate (=recall): $\frac{TP}{P}$.
(P = # positive instances (ground truth))

		Truth (+)	Truth (−)
Pred	(+)	TP 1	FP 1
	(−)	FN 1	TN 2

		Truth (+)	Truth (−)
Pred	(+)	TP 2	FP 0
	(−)	FN 0	TN 3

What are the true positive rates?

Conditional on outcome

True positive rate (=recall): $\frac{TP}{P}$.
(P = # positive instances (ground truth))

		Truth (+)	**Truth** (−)
Pred	(+)	TP 1	FP 1
	(−)	FN 1	TN 2

TPR = 0.5

		Truth (+)	**Truth** (−)
Pred	(+)	TP 2	FP 0
	(−)	FN 0	TN 3

TPR = 1

Conditional on outcome

	group	outcome	prediction
1	A	1	1
2	A	0	0
3	A	1	0
4	B	1	1
5	B	0	0
6	B	1	0
7	B	0	1
...	B	0	1
100	B	0	1

Table: Based on Table 4 from Makhlouf et al., 2021 [link]

Conditional on outcome

	group	outcome	prediction
1	A	1	1
2	A	0	0
3	A	1	0
4	B	1	1
5	B	0	0
6	B	1	0
7	B	0	1
...	B	0	1
100	B	0	1

Table: Based on Table 4 from Makhlouf et al., 2021 [link]

The classifier satisfies equal opportunity.

Conditional on outcome

	group	outcome	prediction
1	A	1	1
2	A	0	0
3	A	1	0
4	B	1	1
5	B	0	0
6	B	1	0
7	B	0	1
...	B	0	1
100	B	0	1

Table: Based on Table 4 from Makhlouf et al., 2021 [link]

The classifier satisfies equal opportunity. However, there are many more false positives in group B. (All B 7—100 are false positives.)

Measuring fairness at the level of groups

Do outcomes systematically differ between different groups?

Three criteria:

equal decision measures	conditional on outcome	conditional on decision
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \| Y$	$Y \perp A \| D$

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on decision

Informally: people with the same decision will have had similar outcomes (regardless of group).

$$Y \perp A|D$$

In a binary classification setting this means $Y \perp A|D = 0$ and $Y \perp A|D = 1$

Individuals are grouped according to the decision, not the actual outcome.

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on decision

First case: $Y \perp A | D = 1$

$$P[Y = 1 | D = 1, A = a] = P[Y = 1 | D = 1, A = b]$$

The precision / PPV (positive predictive value) should be the same for the different subgroups.

This is also called **predictive parity**. Example: When people who are granted loans go on to repay them at the same rate (regardless of the group).

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on decision

Second case: $Y \perp A | D = 0$

$$P[Y = 0 | D = 0, A = a] = P[Y = 0 | D = 0, A = b]$$

Example: All individuals who were denied a loan (D=0) are equally likely to have defaulted if the loan had been granted (Y=0) (regardless of the group).

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on decision

Calibration

- We often have a **score** function R and $D = \mathbb{1}\{R > t\}$
- R is calibrated if $P[Y = 1|R = r] = r$, e.g., 80% of the people with score 0.8 indeed pay back their loan.

R satisfies calibration by group if

$$P[Y = 1|R = r, A = a] = r$$

Calibration by group implies sufficiency.

Measuring fairness at the level of groups

Do outcomes systematically differ between different groups?

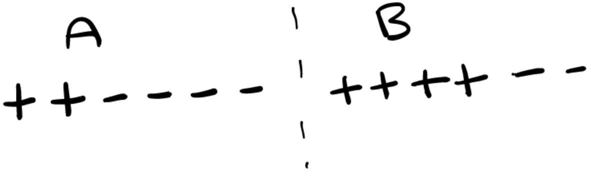
Three criteria:

equal decision measures	**conditional on outcome**	**conditional on decision**
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

Can't we just make systems that satisfy all criteria?

A=sensitive attribute; D=decision; Y=target variable/outcome

We have the following dataset with two groups A and B.
The true labels (+ and –) are shown.



Note: Different base rates (2/6 vs. 4/6).

Is it possible for a classifier to satisfy all criteria?
Remember: statistical parity (equal % of positive predictions), equal of opportunity (equal TPR/recall), predictive parity (equal PPV/precision)

Impossibilities

Bad news! :(

Any 2 of these 3 criteria are mutually exclusive!! (under mild assumptions).

equal decision measures	conditional on outcome	conditional on decision
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

So: We need to make an active choice!
Involve stakeholders and domain experts.

Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big Data, Special issue on Social and Technical Trade-Offs (2017) [link]
Inherent Trade-Offs in the Fair Determination of Risk Scores, Kleinberg et al., Innovations in Theoretical Computer Science (ITCS) 2017 [link]

Impossibilities

Bad news! :(

Any 2 of these 3 criteria are mutually exclusive!! (under mild assumptions).

equal decision measures	**conditional on outcome**	**conditional on decision**
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

So: We need to make an active choice!
Involve stakeholders and domain experts.

Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big Data, Special issue on Social and Technical Trade-Offs (2017) [link]
Inherent Trade-Offs in the Fair Determination of Risk Scores, Kleinberg et al., Innovations in Theoretical Computer Science (ITCS) 2017 [link]

Impossibilities

Suppose we have two groups A and B:

$$FPR_A = \frac{p_A}{1-p_A} \frac{1-PPV_A}{PPV_A}(1-FNR_A)$$

$$FPR_B = \frac{p_B}{1-p_B} \frac{1-PPV_B}{PPV_B}(1-FNR_B)$$

p: prevalence
PPV : positive predictive value (same as precision)
FPR : false positive rates
FNR: false negative rates

See: Chouldechova (2017) [link]

Assumptions:

- the classifier makes mistakes, i.e. FPR_i and $FNR_i > 0$.
- prevalence (base rate) differs between groups, i.e. $p_A \neq p_B$

If PPV is the same across groups (predictive parity), i.e. $PPV_A = PPV_B$, then there's no way to achieve equal FPR and FNR across groups.

COMPAS

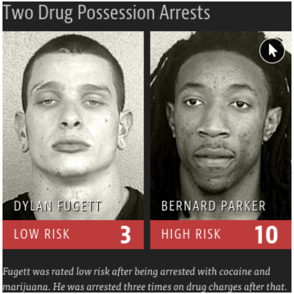


Figure: From ProPublica

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions

Article by ProPublica (Angwin et al., May 23 2016) sparked a lot of debate.

You'll use the COMPAS dataset in the programming exercise.

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

COMPAS

The COMPAS score: risk assessment of recidivism. Used by judges in US.

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%

Figure: From ProPublica

False positive rates and **false negative rates** are not equal!

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

COMPAS

The COMPAS score: risk assessment of recidivism. Used by judges in US.

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%

Figure: From ProPublica

False positive rates and **false negative rates** are not equal!

Response by COMPAS developers (Northpointe): COMPAS satisfies **equal positive predictive** values (Dieterich et al. 2016, [url])

“Bias preserving” vs “bias transforming”

- **Bias preserving**: System should reflect the status quo/training data. Make society not more unequal than it currently is.
  - Quick check: A perfect classifier (zero error according to the labels in the data) satisfies these criteria.
  - Example: Equalized odds, equal opportunity.
  - Focus on *error rates*.
- **Bias transforming**: Acknowledge that the status quo is a result of existing inequalities.
  - Requires making an explicit decision regarding which biases a system should exhibit.
  - Example: Demographic parity.
  - Focus on *decision rates*.

Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, Wachter et al., West Virginia Law Review, 2021 [link]

“Bias preserving” vs “bias transforming”

Wachter et al.: *“By design, bias preserving metrics run the risk of ‘freezing’ or locking in social injustices and discriminatory effects which does not align well with the core aim of EU non-discrimination law: to achieve substantive equality.”*

But:

- Blindly enforcing demographic parity e.g., in lending applications, can make things worse! Individuals may not be able to repay, bankruptcy, etc.
- There are settings where “bias preserving” is suitable, e.g., when we do have an unbiased “ground truth”.

Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, Wachter et al., West Virginia Law Review, 2021 [link]

Which criteria should we use?

Key question: Do we have “ground truth” labels? (If not: statistical parity, conditional statistical parity)

For the following tasks, do we have “ground truth” labels available?
job hiring? college admission? speech recognition?

See also: On the Applicability of Machine Learning Fairness Notions, Makhlouf et. al., ACM SIGKDD Explorations Newsletter 2021 [link]

Broader applications

Note: We have focused on decision making settings, but the same criteria can also be applied to other classification problems (e.g., language identification, part-of-speech tagging, image classification).

Example:
A sentiment classification system that classifies tweets into positive and negative sentiment. We have 2 groups: older and younger Twitter users. We want to use the system to measure public opinion about Dutch politicians.

Is a “bias preserving” or a “bias transforming” criterion more appropriate?

Reflection and outlook

Fairness criteria don’t capture everything! They can’t be “proof” that a system is fair!

Literature

- Chapter 3 "*Classification*" of `https://fairmlbook.org/` "Fairness and machine learning" book, by Solon Barocas, Moritz Hardt, Arvind Narayanan.
    - You can skip: 'Calibration by group as a consequence of unconstrained learning' (19–20) and 'Relationships between criteria' (21–24)
- "*Machine Bias*", Angwin et al., ProPublica, 2016 [link]

Next time

Do the short quiz on Blackboard by **Monday 12pm**
Note: the quizzes are optional.

Start the programming assignment!

Next time:

- We'll continue looking at approaches to measure the fairness of AI systems, focusing on other types of biases (e.g. representations) and limitations of the approaches we discussed today.

Recap:

- vector representations, kNN, linear algebra

Announcement

Paper review preferences are due May 2nd. The week after is a *short* week (Ascension Day).

To give you more time to work on the assignment, if you submit your preferences earlier, we will assign you the papers earlier.

Something to think about

What do you think are the differences between human decision making and AI-(supported) decision making, e.g. in terms of bias, impact, and interventions?