Is hiding fair?

Bettina Berendt

TU Berlin, Weizenbaum Institute for the Networked Society, and KU Leuven

Slides will go on: www.berendt.de/bettina

ACDL 2021

21 July 2021

Who am I, and why am I giving this presentation?

I am professor for Internet and Society at the Faculty of Electrical Engineering and Computer Science at Technische Universität Berlin, Director of the Weizenbaum Institute for the Networked Society, and guest professor in the Declarative Languages and Artificial Intelligence Group DTAI of the Department of Computer Science at KU Leuven.

Overview of Research Areas

Note: Many of these areas overlap - for example, Web mining methods are used to understand and support authors

- General
- AI, data science, and ethics
- Privacy, non-discrimination, fairness, diversity, and related topics
- Web; Social Media; Text, News and Blogs Mining; Learner and Author Support / Knowledge Management
- Information Search and Ubiquitous Information
- Semantic Web Mining, Ontologies and Knowledge Discovery
- Web Usage Mining, Query Mining
- Other data mining topics
- e-Commerce, Web Metrics, Evaluation of Information Systems

What is my goal with these two lectures?

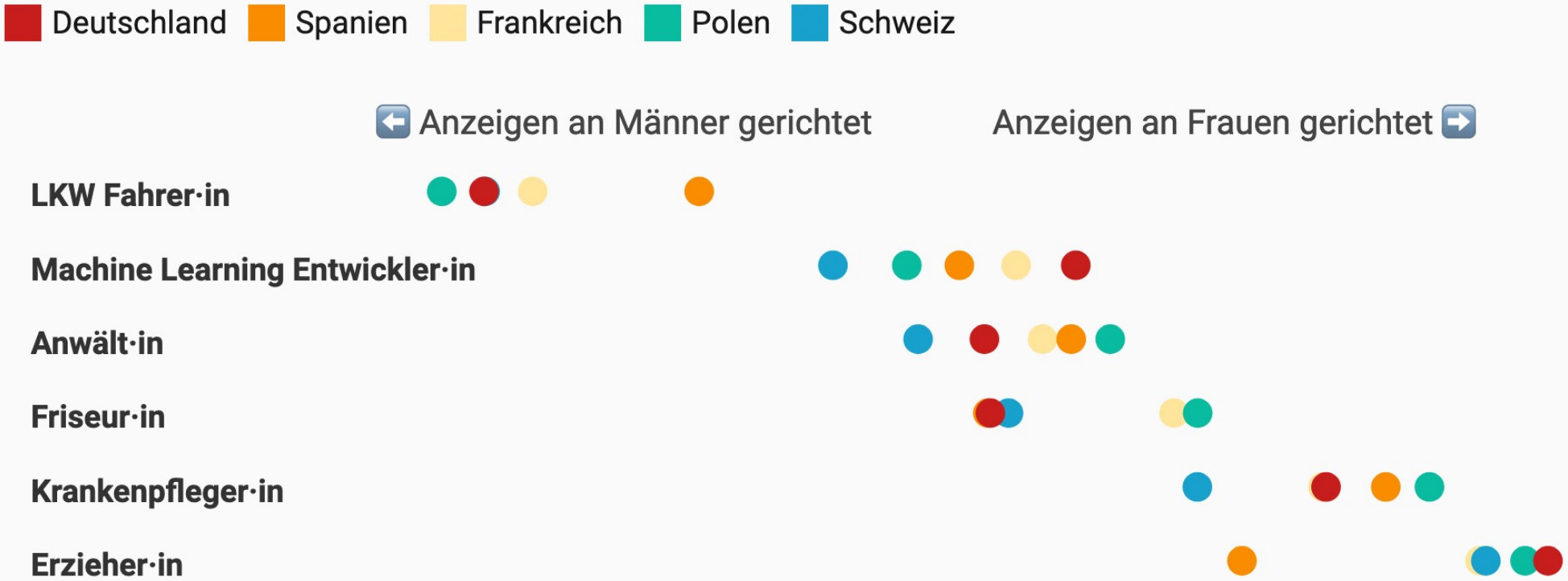
- A LOT has been and is being said and written about these and related problems.
- Often, computer scientists come across in these accounts as either “the villains” or ”the fixers” or “the clueless”
- Interdisciplinary collaboration, and collaboration with non-academics, are definitely needed to address unfairness and discrimination
- And we can contribute insights and actions that only we can contribute
- I hope to convince you of the relevance and intellectual satisfaction of doing so!

Is this fair?

Facebook verwendet grobe Stereotypen, um Anzeigenschaltung zu optimieren

"Facebook uses rough stereotypes to optimize the selection/presentation of advertising"

Wir haben Anzeigen für sechs verschiedene Stellenangebote in fünf Ländern gekauft. So optimierte Facebook die Anzeigenimpressionen, aufgeschlüsselt nach Geschlecht.



Die Grafik basiert auf 102.472 Anzeigenimpressionen zwischen 27.9. und 3.9.

Chart: AlgorithmWatch • Get the data • Created with Datawrapper

https://algorithmwatch.org/de/automatisierte-diskriminierung-facebook-verwendet-grobe-stereotypen-um-die-anzeigenschaltung-zu-optimieren/

In February 2019, Nijeer Parks was accused of shoplifting candy and trying to hit a police officer with a car at a Hampton Inn in Woodbridge, N.J. The police had identified him using facial recognition software, even though he was 30 miles away at the time of the incident.

Mr. Parks spent 10 days in jail and paid around $5,000 to defend himself. In November 2019, the case was dismissed for lack of evidence.

Mr. Parks, 33, is now suing the police, the prosecutor and the City of Woodbridge for false arrest, false imprisonment and violation of his civil rights.

He is the third person known to be falsely arrested based on a bad facial recognition match. In all three cases, the people mistakenly identified by the technology have been Black men.

Amid the controversy, the Detroit City Council last fall approved a nearly $200,000 facial recognition technology contract with South Carolina-based DataWorks Plus, which funds software maintenance and equipment support. The contract expires Sept. 30, 2022.

Mr. Parks spent 10 days in jail and paid around $5,000 to defend himself. In November 2019, the case was dismissed for lack of evidence.

Mr. Parks, 33, is now suing the police, the prosecutor and the City of Woodbridge for false arrest, false imprisonment and violation of his civil rights.

He is the third person known to be falsely arrested based on a bad facial recognition match. In all three cases, the people mistakenly identified by the technology have been Black men.

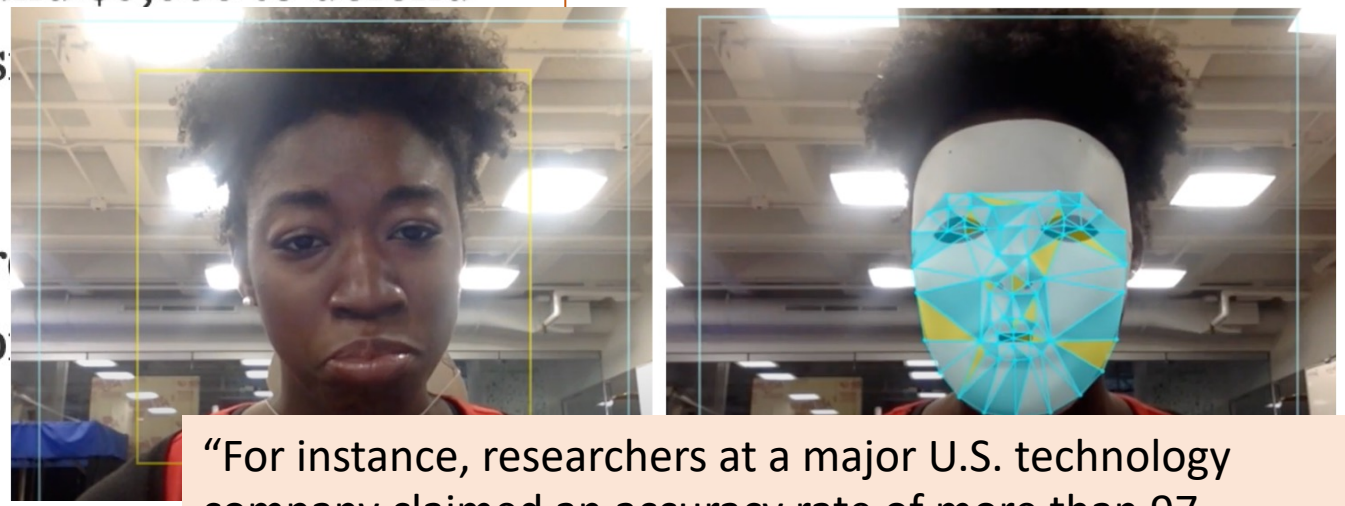
https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html
https://eu.detroitnews.com/story/news/politics/2021/07/13/house-panel-hear-michigan-man-wrongfully-accused-facial-recognition/7948908002/
https://cdn.kastatic.org/ka-perseus-images/6a738b97d70ebd4b0fc4df4124d2dbf7b23b336f.png
https://www.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002/

In February 2019, Nijeer Parks was accused of shoplifting candy and trying to hit a police officer with a car at a Hampton Inn in Woodbridge, N.J. The police had identified him using facial recognition software, even though he was 30 miles away at the time of the incident.

Mr. Parks spent 10 days in jail and paid around $5,000 to defend himself. In November 2019, the case was dis
evidence.

Mr. Parks, 33, is now suing the police, the pr
of Woodbridge for false arrest, false impriso
his civil rights.

He is the third person known to be falsely arrested
facial recognition match. In all three cases, the peop
identified by the technology have been Black men.



"For instance, researchers at a major U.S. technology company claimed an accuracy rate of more than 97 percent for a face-recognition system they'd designed. But the data set used to assess its performance was more than 77 percent male and more than 83 percent white." (Buolamwini & Gebru, 2018; citation from MIT News)

Non-discrimination: a fundamental right
(and a definition of fairness via non-discrimination)

Article 7, Universal Declaration of Human Rights

All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination.

Article 21, European Charter of Fundamental Rights: Non-discrimination

1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

2. Within the scope of application of the Treaty establishing the European Community and of the Treaty on European Union, and without prejudice to the special provisions of those Treaties, any discrimination on grounds of nationality shall be prohibited.

"**A system discriminates unfairly** if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate."

Friedman and Nissenbaum (1996) Bias in Computer Systems

Direct and indirect discrimination (~ disparate treatment and disparate impact)

Direct discrimination [occurs] when you are treated worse than another person or other people because:

- you have a protected characteristic
- someone thinks you have that protected characteristic (known as discrimination by perception)
- you are connected to someone with that protected characteristic (known as discrimination by association)

Your circumstances must be similar enough to the circumstances of the person being treated better […].

If you cannot point to another person who has been treated better, it is still direct discrimination if you can show that a person who did not have your protected characteristic would have been treated better in similar circumstances.

Indirect discrimination happens when there is a policy that applies in the same way for everybody but disadvantages a group of people who share a protected characteristic, and you are disadvantaged as part of this group. […]

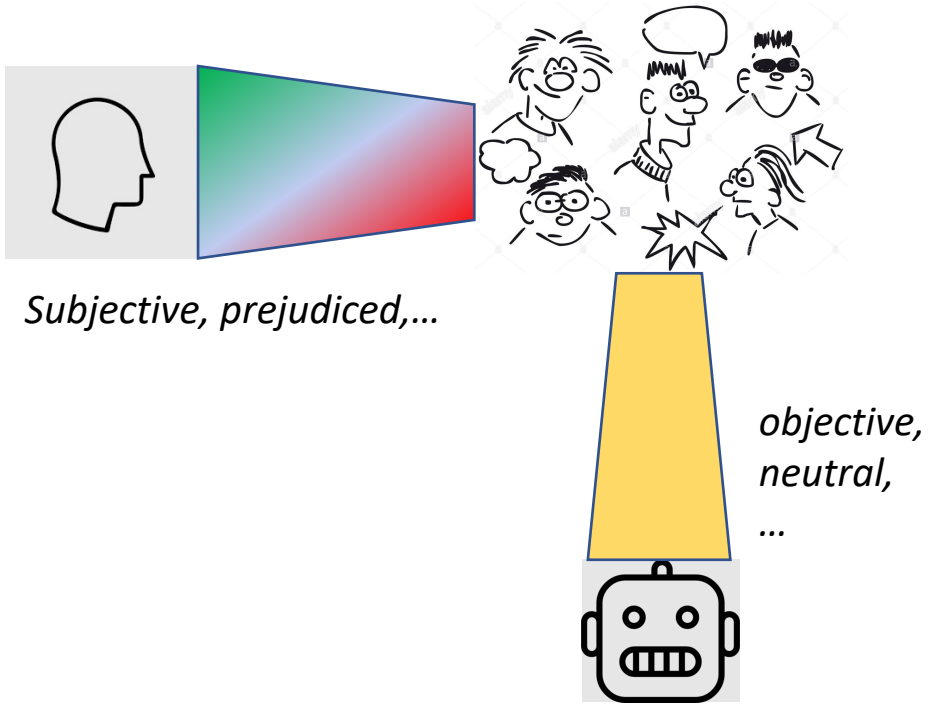
A 'policy' can include a practice, a rule or an arrangement.

It makes no difference whether anyone intended the policy to disadvantage you or not. […]

If the organisation can show there is a good reason for its policy, it is not indirect discrimination. This is known as objective justification.

https://www.equalityhumanrights.com/en/advice-and-guidance/what-direct-and-indirect-discrimination

Why do we hope that AI / ML / ADMS will make better decisions than humans in the first place?



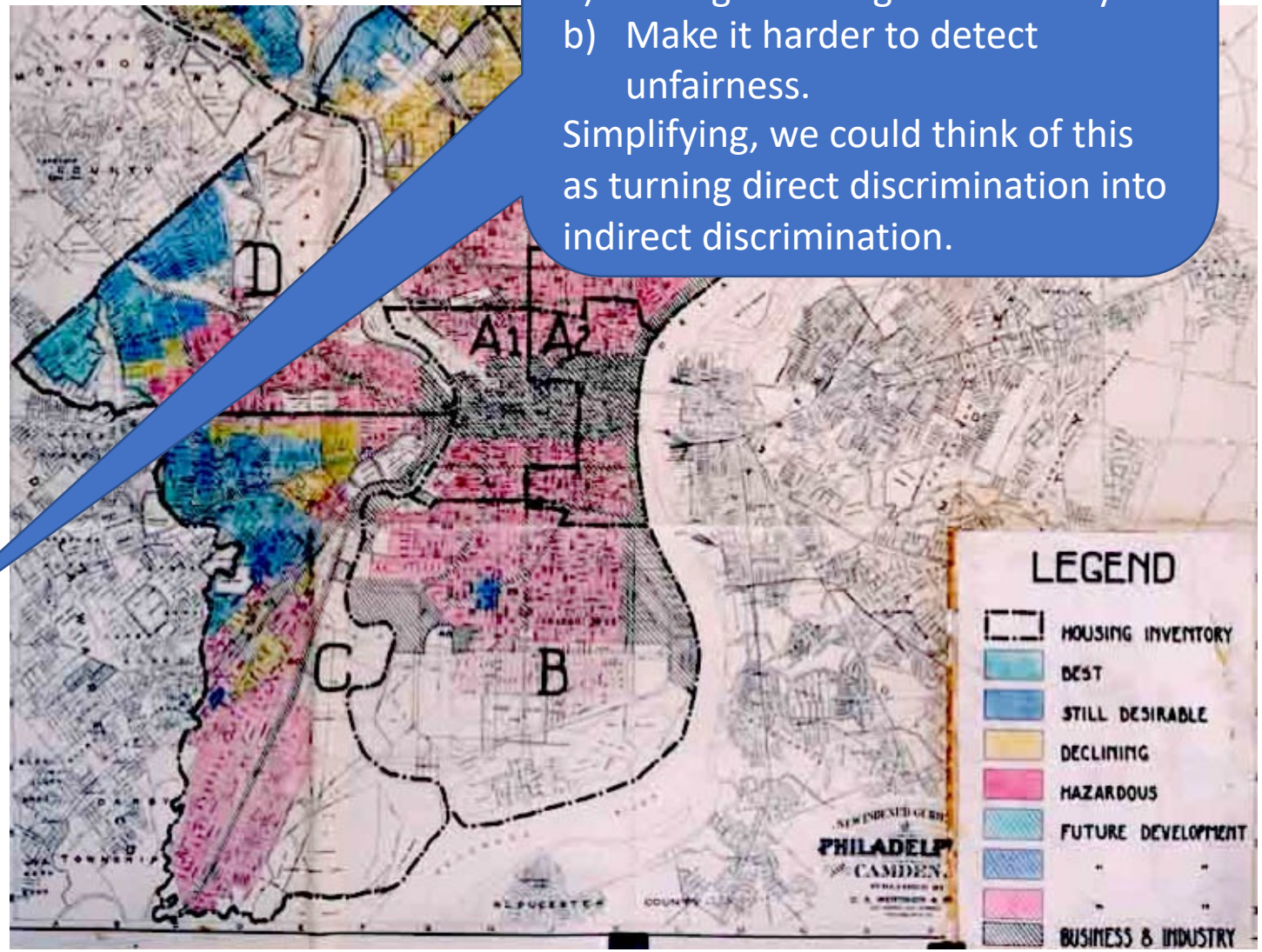
Is hiding fair? First try.

- Classic example: Black people in the US have more difficulty obtaining a loan than white people. Why not just omit the *protected attribute* from the data?
- ”Redlining” – ZIP code being highly correlated with so-called ‘race’ creates indirect discrimination

“Stupid” hiding may cause an ML model to learn to predict from proxies and thus

a) Change nothing substantially
b) Make it harder to detect unfairness.

Simplifying, we could think of this as turning direct discrimination into indirect discrimination.



https://upload.wikimedia.org/wikipedia/commons/4/41/Home_Owners%27_Loan_Corporation_Philadelphia_redlining_map.jpg

Table 3 Learned models.

	salary				rank		degree		years		gender
Standard model (M1):	w	$=$	11956	$+$	$4993r$	$+$	$398d$	$+$	$103y$	$-$	$950s$
Blind model (M0):	w	$=$	11604	$+$	$5231r$	$+$	$179d$	$+$	$88y$		
Only males (MM):	w	$=$	11705	$+$	$5032r$	$-$	$31d$	$+$	$129y$		
Only females (MF):	w	$=$	10117	$+$	$4567r$	$+$	$2399d$	$+$	$116y$		

"Stupid" hiding may aggravate unfairness.

In most realistic cases, not only removing the sensitive variable does not make regression models fair, but on the contrary, such a strategy is likely **to amplify discrimination.**

I. Zliobaite, B. Custers. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artif. Intell. Law 24(2): 183-201, 2016.

https://pair.withgoogle.com/explorables/hidden-bias/

Slide from Salvatore Ruggieri. *Discrimination and Fairness*. NoBIAS OnBoardingWeek, 23 March 2021 (added: the callout)

A short summary of data-privacy metrics (1): from k-anonymity to t-closeness

3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Homogeneity attack

Bob

Zipcode	***Age***
47678	27

Background knowledge attack

Carl

Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Slide from Vitaly Shmatikov (2009). k-Anonymity and other Cluster-Based Methods. https://www.cs.utexas.edu/~shmat/courses/cs380s_fall09/21kanon.ppt

l-Diversity

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class

Neither Necessary, Nor Sufficient

Original dataset

…	Cancer
…	Cancer
…	Cancer
…	Flu
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Flu
…	Flu

99% have cancer

Anonymization A

Q1	Flu
Q1	Flu
Q1	Cancer
Q1	Flu
Q1	Cancer
Q1	Cancer
Q2	Cancer

Anonymization B

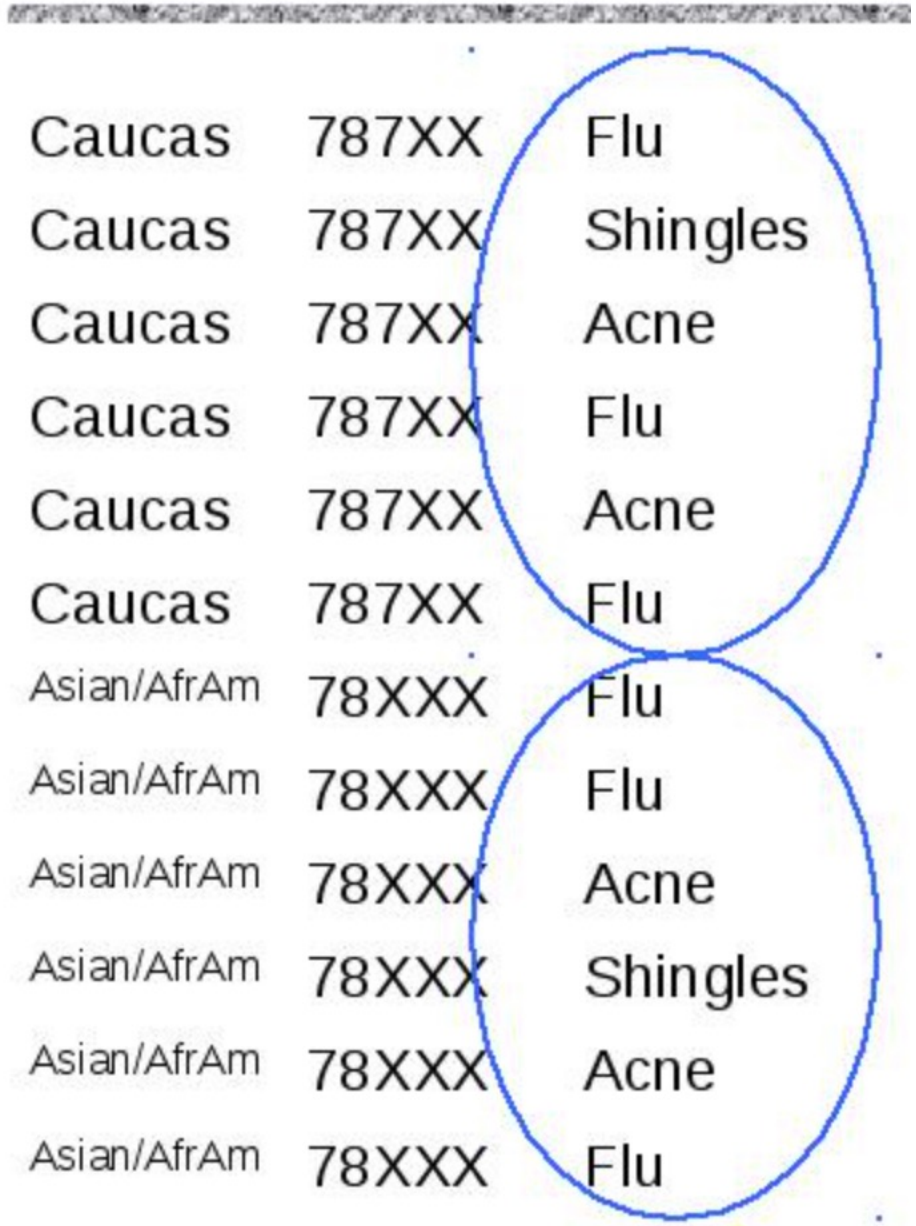
Q1	Flu
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q2	Flu

99% cancer ⇒ quasi-identifier group is not "diverse"
…yet anonymized database does not leak anything

50% cancer ⇒ quasi-identifier group is "diverse"
This leaks a ton of information

Slide from Vitaly Shmatikov (2009). k-Anonymity and other Cluster-Based Methods. https://www.cs.utexas.edu/~shmat/courses/cs380s_fall09/21kanon.ppt

t-Closeness



Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database

Slide from Vitaly Shmatikov (2009). k-Anonymity and other Cluster-Based Methods. https://www.cs.utexas.edu/~shmat/courses/cs380s_fall09/21kanon.ppt

Data privacy via t-closeness – high-level comparison with – Non-discrimination

-----Quasi-identifier------		Sensitive attribute
Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database

Similarities?
Differences?

-Protected attribute-		Decision attribute
Caucas	(	No
Caucas	(	Yes
Caucas	(	Yes
Caucas	(	Yes
Caucas	(	No
Caucas	(	No
Asian/AfrAm	(	No
Asian/AfrAm	(	No
Asian/AfrAm	(	Yes
Asian/AfrAm	(	No
Asian/AfrAm	(	Yes
Asian/AfrAm	(	No

Other attributes, considered not discriminatory

Of course, making a table t-close (or “fair”) is not straightforward – Example: 2 data pre-processing strategy from Ruggieri (2014)

“Smart” hiding may reduce unfairness in the sense of “different distributions“.

Sample dataset

ID	purpose	emp	sex	decision
1	housing	no	female	-
2	housing	no	female	-
3	housing	no	female	+
4	housing	no	male	-
5	housing	no	male	+
6	housing	yes	female	-
7	housing	yes	female	+
8	housing	yes	female	+
9	housing	yes	male	-
10	housing	yes	male	-
11	housing	yes	male	+
12	housing	yes	male	+
13	car	no	female	+
14	car	no	male	-
15	car	no	male	+
16	car	yes	female	-
17	car	yes	male	+

Output of dMondrian

ID	purpose	emp	sex	decision
1	housing-car	no	female	-
2	housing-car	no	female	-
3	housing-car	no	female	+
13	housing-car	no	female	+
4	housing-car	no	male	-
14	housing-car	no	male	-
5	housing-car	no	male	+
15	housing-car	no	male	+
6	housing-car	yes	female	-
16	housing-car	yes	female	-
7	housing-car	yes	female	+
8	housing-car	yes	female	+
9	housing-car	yes	male	-
10	housing-car	yes	male	-
11	housing-car	yes	male	+
12	housing-car	yes	male	+
17	housing-car	yes	male	+

Figure 3: Sample dataset and *dMondrian* output for $t = 0.25$.

Output of dSabre

ID	purpose	emp	sex	decision
1	housing	no-yes	female	-
3	housing	no-yes	female	+
4	housing	no-yes	male	-
5	housing	no-yes	male	+
11	housing	no-yes	male	+
6	housing	yes	female	-
7	housing	yes	female	+
9	housing	yes	male	-
12	housing	yes	male	+
2	housing-car	no	female	-
13	housing-car	no	female	+
14	car	no	male	-
15	car	no	male	+
16	housing-car	yes	female	-
8	housing-car	yes	female	+
10	housing-car	yes	male	-
17	housing-car	yes	male	+

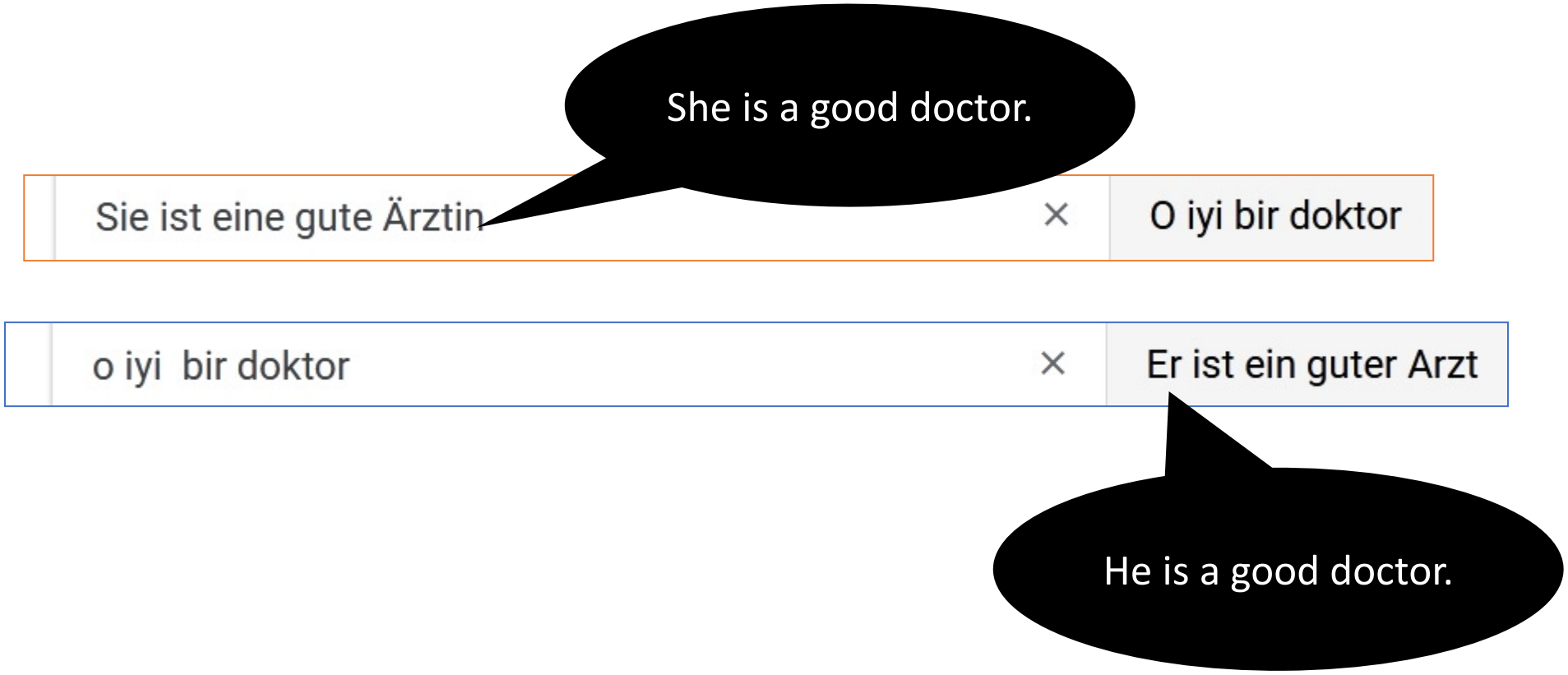
Side note: The GDPR recognises this problem

Recital 71 *(complemented by other Recitals)*

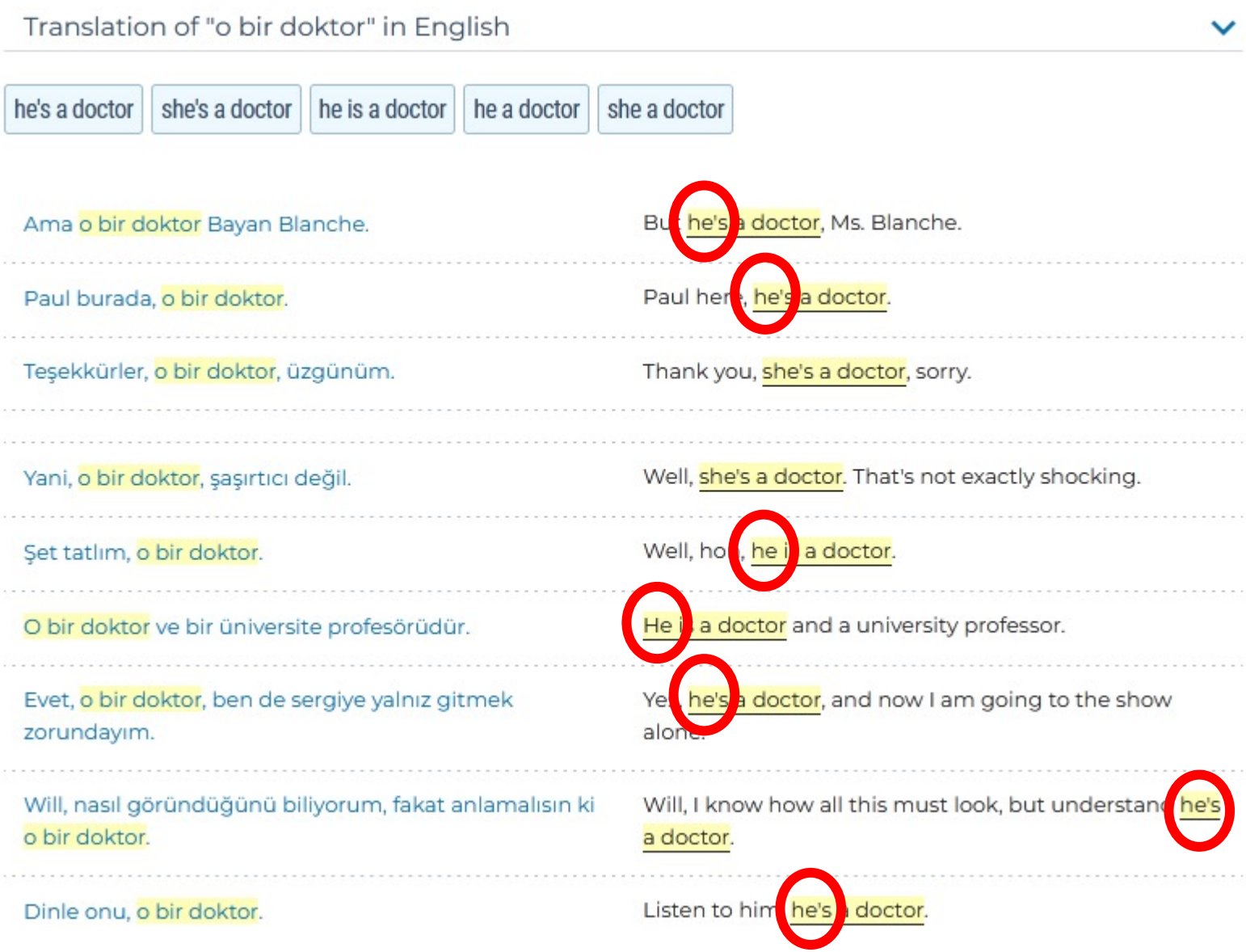
In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, 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 that result in measures having such an effect. Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.

So where does the unfairness/discrimination come from?

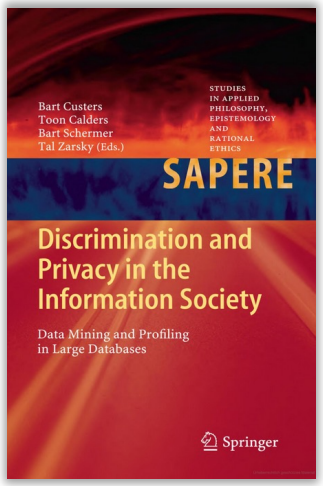
Ex. 1: AI language processing
Bias in machine translation



Why? Machine learning from data that reflects historical & current inequalities



Countermeasures: Modifying data and algorithms; collaborating



FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

Full book as PDF

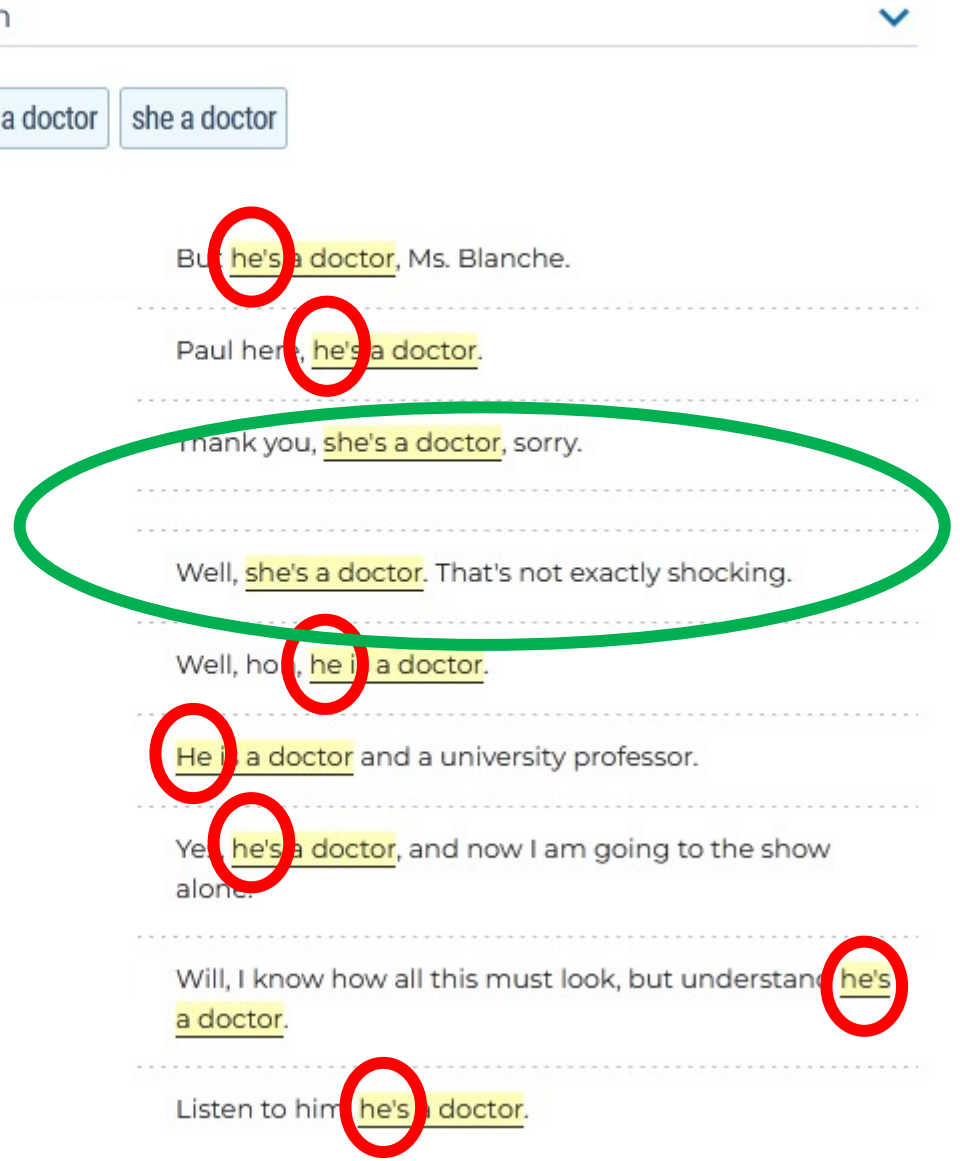
The New York Times | https://www.nytimes.com/2021

Using A.I. to Find Bias in A.I.

The problem of bias in artificial intelligence is facing increasing scrutiny from regulators and is a growing business for start-ups and tech stalwarts.

By Cade Metz

& & & …



Impact of algorithm & problem formulation – Countermeasures – Limitations

Why is this a problem?

- “representational harm”
  - Prejudiced language and its effects
- “allocational harm”
  - Derived decisions, for example
  - Prediction of interests & “optimized” advertising (see introductory example)
  - Members of protected groups may see the ad, but not apply for it
  - Prediction of job skills & further processing (or not) of applications
  - Prediction of job chances & access to support based on this (e.g. training)
    - Note: The system I will describe under this heading does not operate on language data. Basic problems of machine learning remain.

Effects of “neutrally” formulated job ads

- Typical “agent-oriented” words to describe skills discourage women from applying for the job
- Rephrasing using more “communal-oriented” words significantly increased these percentages
- No such effect for men – they apply anyway.

Agent-oriented	Communal-oriented
Independent	Taking responsibility
Ambitious, competitive	Committed, engaged
Goal-oriented	reliable
Direct, determined	honest

Burell (2021). Die Diskriminierung steckt oft im Detail. https://www.spiegel.de/start/stellenanzeigen-werden-oft-fuer-maenner-formuliert-wie-frauen-trotzdem-den-job-bekommen-a-2ff0215c-009c-48b1-b045-a462ca808cd7

Ex. 2: Algorithms used in recruitment: indirect discrimination – based on biased data

- *The [Amazon] team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters. […] In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges,*
- *Amazon edited the programs to make them neutral to these particular terms. But that was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory, the people said.*
- *The Seattle company ultimately disbanded [the algorithm]. *)*

- Most of the media coverage at the time focused on the pitfall of using historical data to train algorithms.
- But …

*) https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Ex. 2: Algorithms used in recruitment: indirect discrimination – based on biased data – or is it?

- “As an AI practitioner myself, and I’m more interested in the technical and business details. I was frustrated by how shallow — and sometimes wrong — most of the reporting was.
- It’s essential to realize that the main reason the piece got some much publicity is the notoriety of Amazon
- Algorithms aren’t morally biased. […] Even data isn’t biased. […] It means that to understand what happened at Amazon — or in any AI project — **we need to understand the human designers, their goals, the resources they had, and the choices they had to make.”**

Lauret, J. 2019. Amazon’s sexist AI recruiting tool: how did it go so wrong? https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e

Ex. 2: Algorithms used in recruitment: indirect discrimination – based on biased data – or is it?

1. **Business objective:** Resume screening is an input to a complex hiring process.
   - Increasing diversity of the workforce -> minimise false negatives.
   - BUT cost of hiring the wrong person is high -> minimise false positives.
2. **Data:** it seems that Amazon only used the resumes submitted by past candidates. There is self-selection bias. All the CVs come from people who already think that Amazon would be a good fit for them.
   - Ok to use these as training data if the goal is to only screen applications. But could be a problem if the goal is to scrape the Web (e.g. LinkedIn).
   - What if there are no female-applicant data?
3. **Target variable and cost function:**
   - Target variable could be: invited to interview? Invited to follow-up interview? Good job performance? -> Which issues do these pose?
   - Cost function? (E.g. MAE – 1 vs. 3 or 5 vs. 3.5 stars of 5 – these errors can be vastly different)

What if there is "nothing to hide"? (Or is there?)

Lauret, J. 2019. Amazon's sexist AI recruiting tool: how did it go so wrong?
https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e

Ex. 3: The AMS algorithm

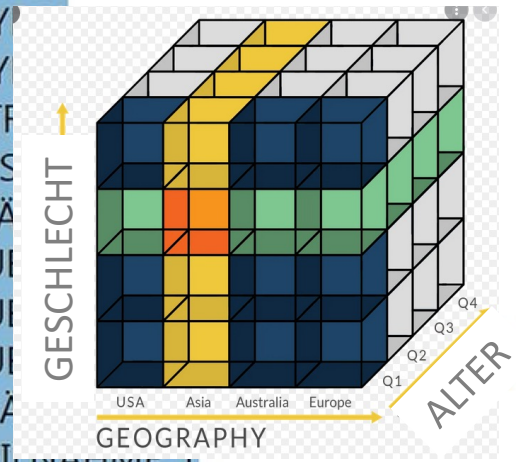
- Project of the Austrian Public Employment Service (AMS)
- The “Job market chances system” is supposed to predict future chances of jobseekers on the basis of statistics from past years.
- Jobseekers are classified into three groups according to their “integration chance”.
- The „middle“ segment is the focus for allocating support (esp. Training measures).
- Sociotechnical analysis: Allhutter et al. (2020) – Conclusions:
  - The IC value has pervasive and broad consequences for AMS consultancy practice and for jobseekers.
  - Transparency and rights of objection as well as public participation are therefore required.

- Current status: AMS vs. data protection authority – Federal Administrative Court

GENDER_FEMALE

CARE OBLIGATIO

BE_INT
1 = f (0,10
2 – 0,14 x GESCHLECHT_WEIBLICH
– 0,13 x ALTERSGRUPPE_30_49
– 0,70 x ALTERSGRUPPE_50_PLUS
+ 0,16 x STAATENGRUPPE_EU
– 0,05 x STAATENGRUPPE_DRITT
3 + 0,28 x AUSBILDUNG_LE
+ 0,01 x AUSBILDUNG_MAX_PLUS
4 – 0,15 x BETREUUNGSPFLICHTIG
– 0,34 x RGS_TYP_2
– 0,18 x RGS_TYP_3
5 – 0,83 x RGS_TY
– 0,82 x RGS_TY
6 – 0,67 x BEEINTR
+ 0,17 x BERUFS
– 0,74 x BESCHÄ
+ 0,65 x FREQUE
+ 1,19 x FREQUE
+ 1,98 x FREQUE
7 – 0,80 x GESCHÄ
– 0,57 x MN_TEILNAHME_1
– 0,21 x MN_TEILNAHME_2
– 0,43 x MN_TEILNAHME_3)

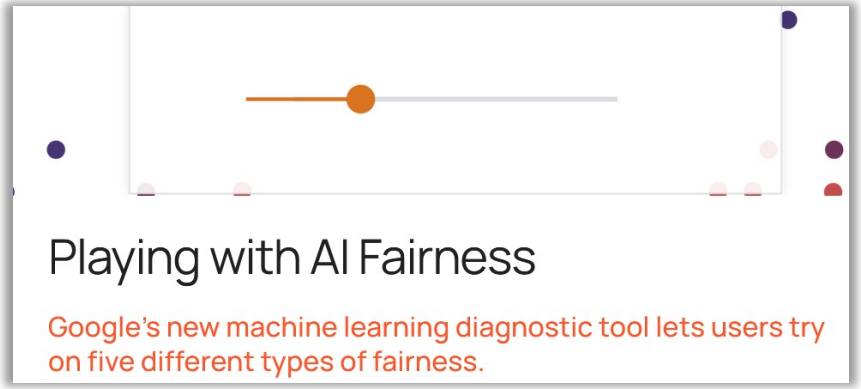


How could such a system be “non-discriminatory”?

- Should (e.g.) men and women be classified into the middle segment equally often?
  - “equality of outcome”
- Should error rates be the same across these groups?
  - E.g. for the desired class: “equality of opportunity”
- Should we compare on the group level or on the level of individuals who are similar except in the protected attribute?
  - “group fairness / individual fairness”

➔ Debate about the metrics for “(non)discrimination” resp. “fairness”

- Note: this relates to millenia-old debates about equality, justice, fairness, etc.!



https://pair-code.github.io/what-if-tool/ai-fairness.html

Side remark: Analysis without data&algorithm?

- Allhutter et al. had no access to the training data or current model.
- In the OLAP-like model, it might be more difficult to detect discrimination than in the ”leaked” prior (regression) model.
- Also, it may not be direct discrimination as in the regression model.
- But even based on this analysis only, they asked pertinent questions.
- (Side note: This shows the value of sociotechnical analysis!)
- Still, for us as computer scientists it is often clearer to argue from a position of having access to data and/or algorithms.
- For this, we will go back to examples where also data have been analysed.
- We then return to the AMS system

COMPAS: Indirect discrimination? (White/black was not a feature in the model)

Prediction Fails Differently for Black Defendants

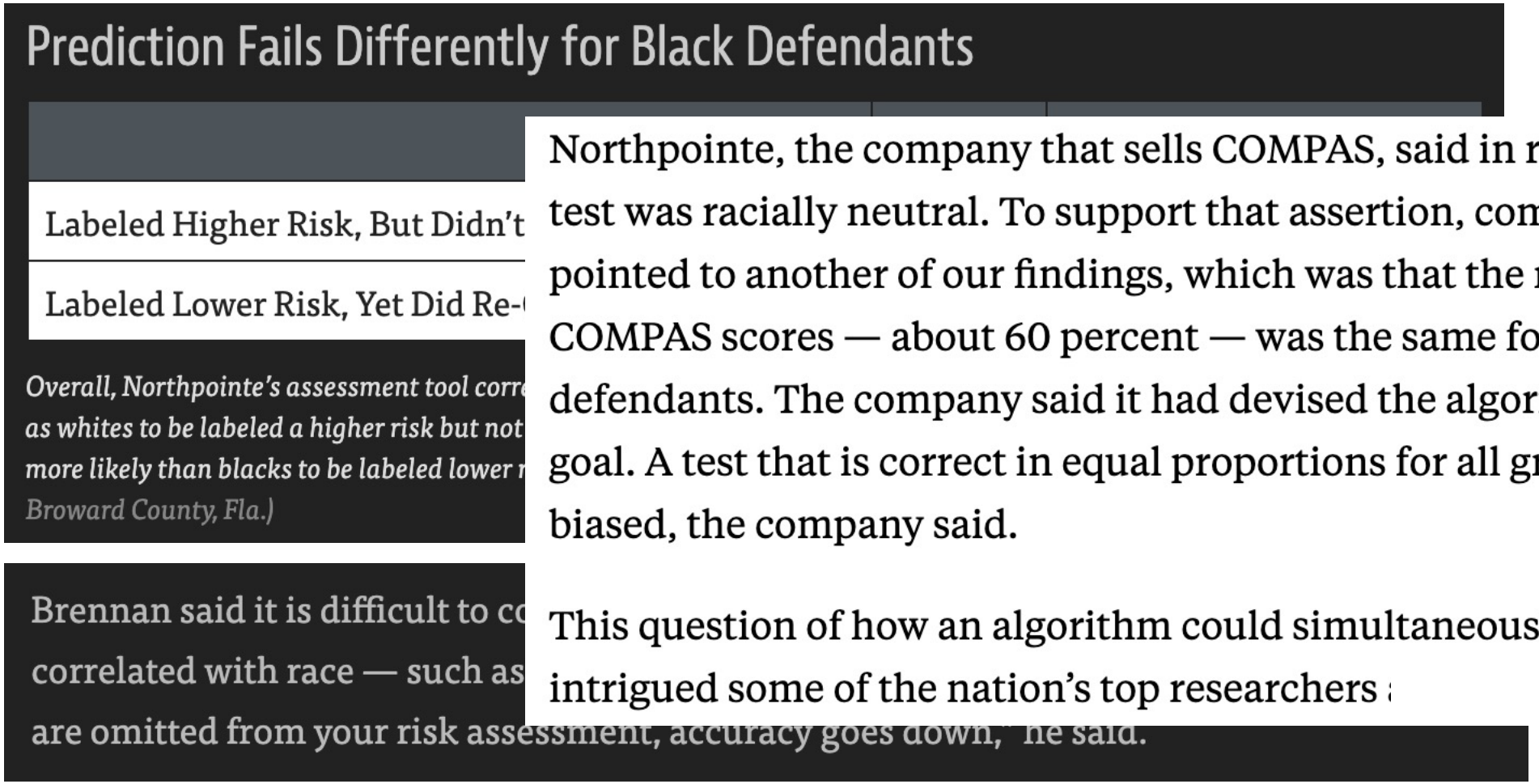
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Brennan said it is difficult to construct a score that doesn't include items that can be correlated with race — such as poverty, joblessness and social marginalization. "If those are omitted from your risk assessment, accuracy goes down," he said.

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

COMPAS: No discrimination?



Prediction Fails Differently for Black Defendants

Labeled Higher Risk, But Didn't	
Labeled Lower Risk, Yet Did Re-	

Overall, Northpointe's assessment tool corre
as whites to be labeled a higher risk but not
more likely than blacks to be labeled lower r
Broward County, Fla.)

Brennan said it is difficult to co
correlated with race — such as
are omitted from your risk assessment, accuracy goes down," he said.

Northpointe, the company that sells COMPAS, said in response that the test was racially neutral. To support that assertion, company officials pointed to another of our findings, which was that the rate of accuracy for COMPAS scores — about 60 percent — was the same for black and white defendants. The company said it had devised the algorithm to achieve this goal. A test that is correct in equal proportions for all groups cannot be biased, the company said.

This question of how an algorithm could simultaneously be fair and unfair intrigued some of the nation's top researchers

https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say

Examples here: Caton & Haas (2020). Hutchinson & Mitchell (2019)

3.2.2 Confusion Matrix-based Metrics

While parity-based metrics typically consider variants of the predicted positive rate $Pr(\hat{y} = 1)$, confusion matrix-based metrics take into consideration additional aspects such as True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR). The advantage of these types of metrics is that they are able to include underlying differences between groups who would otherwise not be included in the parity-based approaches. This is related to the Separation criterion that was defined in subsection 3.1.

Equal Opportunity: As parity and disparate impact do not consider potential differences in groups that are being compared, [129, 223] consider additional metrics that make use of the FPR and TPR between groups. Specifically, an algorithm is considered to be fair under equal opportunity if its TPR is the same across different groups.

$$Pr(\hat{y} = 1|y = 1\&g_k) = Pr(\hat{y} = 1|y = 1\&g_j) \quad (6)$$

Equalized Odds (Conditional procedure accuracy equality [27]): Similarly to equal opportunity, in addition to TPR equalized odds simultaneously considers FPR as well, i.e., the percentage of actual negatives that are predicted as positive.

$$Pr(\hat{y} = 1|y = 1\&g_k) = Pr(\hat{y} = 1|y = 1\&g_j) \ \& \ Pr(\hat{y} = 1|y = 0\&g_k) = Pr(\hat{y} = 1|y = 0\&g_j) \quad (7)$$

Overall accuracy equality [27]: Accuracy, i.e., the percentage of overa… ative), is one of the most widely used classification metrics. [27] adjusts… rates across different groups. If two groups have the same accuracy, they…

$$Pr(\hat{y} = 0|y = 0\&g_k) + Pr(\hat{y} = 1|y = 1\&g_k) = Pr(\hat{y} = 0|y…$$

Conditional use accuracy equality [27]: As an adaptation of the overall… procedure and conditional use accuracy do not look at the overall accurac… and negative predictive values.

$$Pr(y = 1|\hat{y} = 1\&g_k) = Pr(y = 1|\hat{y} = 1\&g_j) \ \& \ Pr(y = 0|\hat{y} =$$

Treatment equality [27]: Treatment equality considers the ratio of False… Predictions.

$$\frac{Pr(\hat{y} = 1|y = 0\&g_k)}{Pr(\hat{y} = 0|y = 1\&g_k)} = \frac{Pr(\hat{y} = 1|y}{Pr(\hat{y} = 0|y}$$

Equalizing disincentives [148]: The Equalizing disincentives metric c… and FPR, across the groups and is specified as:

$$Pr(\hat{y} = 1|y = 1\&g_k) - Pr(\hat{y} = 1|y = 0\&g_k) = Pr(\hat{y} = 1|y…$$

Conditional Equal Opportunity [30]: As some metrics can be depend… provide an additional metric that specifies equal opportunity on a specific… τ is a threshold value:

$$Pr(\hat{y} \geq \tau|g_k\&y < \tau\&A = a) = Pr(\hat{y} =\geq \tau|…$$

3.2.3 Calibration-based Metrics

In comparison to the previous metrics which are defined based on the predicted and actual values, calibration-based metrics take the predicted probability, or score, into account. This is related to the Sufficiency criterion that was defined in Section 3.1.

Test fairness/ calibration / matching conditional frequencies ([66], [129]): Essentially, test fairness or calibration wants to guarantee that the probability of $y = 1$ is the same given a particular score. I.e., when two people from different groups get the same predicted score, they should have the same probability of belonging to $y = 1$.

$$Pr(y = 1|S = s\&g_k) = Pr(y = 1|S = s\&g_j) \quad (13)$$

Well calibration [168]: An extension of regular calibration where the probability for being in the positive class also has to equal the particular score.

$$Pr(y = 1|S = s\&g_k) = Pr(y = 1|S = s\&g_j) = s \quad (14)$$

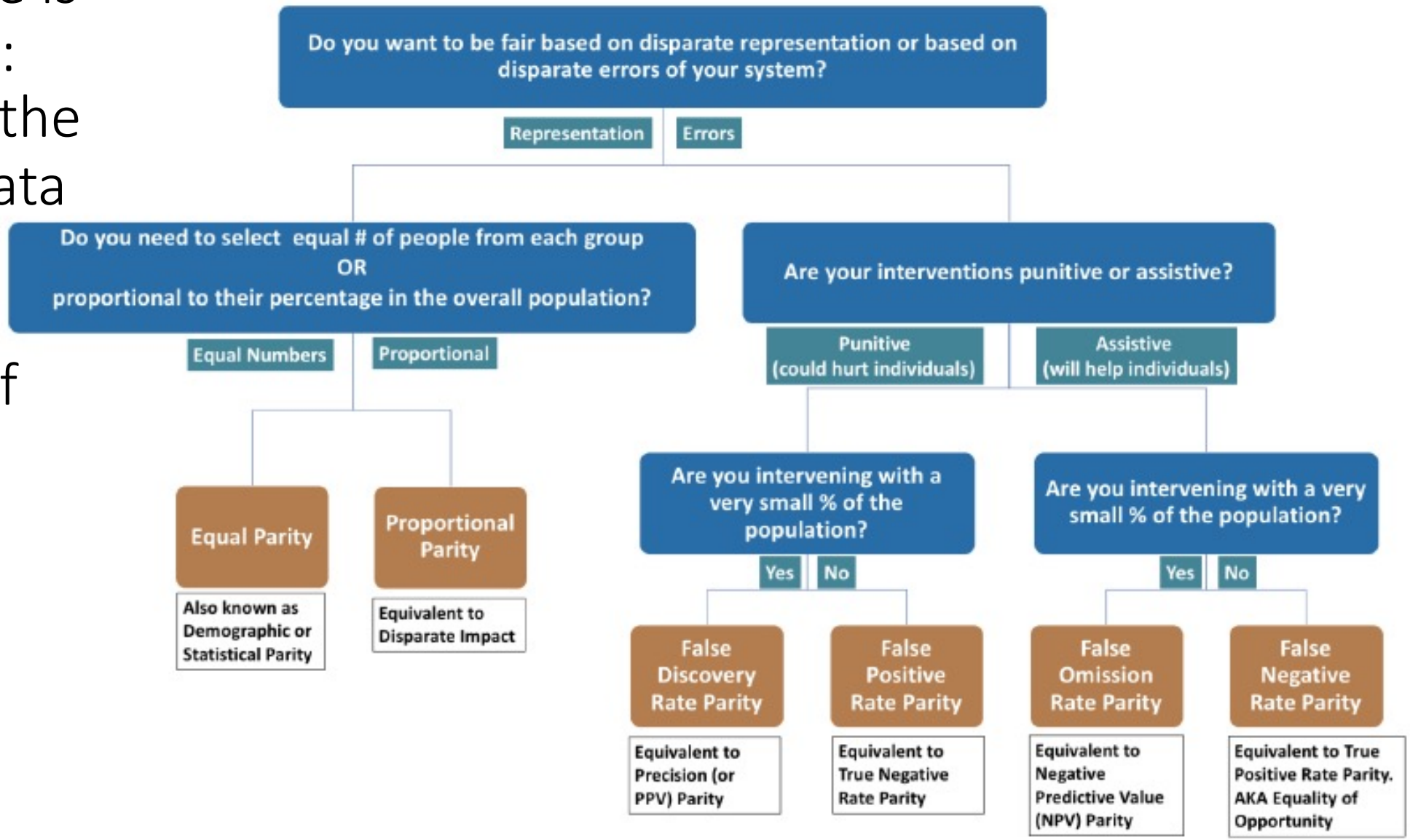
50 Years of Test (Un)fairness: Lessons for Machine Learning FAT* '19, January 29–31, 2019, Atlanta, GA, USA

Source	Criterion	Category	Proposition
Guion (1966)	"people with equal probabilities of success on the job have equal probabilities of being hired for the job"	INDIVIDUAL	Is the use of the test fair?
Cleary (1966)	"a subgroup does not have consistent errors"	NON-COMPARATIVE	Is the test fair to subgroup a?
Einhorn and Bass (1971)†	$Prob(Y > y* \mid R = r_{a*}, A = a)$ is constant for all…	SUBGROUP PARITY	Is the use of the test fair…
			?
at position n†	top-n candidates ranked by model score as it has in the top-n candidates ranked by Y		subgroup a?
Peterson & Novick (1976) *conditional probability and its converse*	$Prob(R >= r_{a*} \mid Y >= y*, A = a)$ is constant for all subgroups a, and $Prob(R < r_{a*} \mid Y < y*, A = a)$ is constant for all subgroups a	SUBGROUP PARITY	Is the use of the test fair with respect to A?
Peterson & Novick (1976) *equal probability and its converse*	$Prob(Y >= y* \mid R >= r_{a*}, A = a)$ is constant for all subgroups a, and $Prob(Y < y* \mid R < r_{a*}, A = a)$ is constant for all subgroups a	SUBGROUP PARITY	Is the use of the test fair with respect to A?

Table 2: Early technical definitions of fairness in educational and employment testing. Variables: R is the test score; Y is the target variable; A is the demographic variable. The Proposition column indicates whether fairness is considered a property of the way in which a test is used, or of the test itself. † indicates that the criterion is discussed in the appendix.

Many metrics, some of which cannot be achieved simultaneously … and the problem isn't exactly new
(more in Isabel Valera's lecture on Friday!)

Metric choice is not arbitrary: A view from the Center for Data Science and Public Policy (University of Chicago)



http://aequitas.dssg.io

Metric choice is not arbitrary: A view from political philosophy / ethics

- We consider different types of equality appropriate / “just” / “fair” in different types of contexts
- E.g. voting: everyone should vote, regardless of talent etc. → equality of outcome
  - tests are not considered acceptable (by most people)
- E.g. economic and social benefits: “We may consider it fair, other things being equal, that the most qualified applicant obtains [a job], and that the most industrious and/or talented individuals deserve more economic benefits than others
  - even if we believe that current systems do not actually ensure a level playing eld, and some level of income redistribution is also morally required”
  - Tests are accepted / desired / required (by most people)
- *(For thoughts about the COMPAS example, see Binns’ pape*r)

Binns (2018)

Metric choice is not arbitrary: A legal viewpoint

Fairness metric	Bias preserving?
1. Group fairness, Statistical (demographic) parity	✗
2. Conditional statistical (demographic) parity, Conditional independence	✗
3. Predictive parity, outcome test	✓
4. False positive error rate balance	✓
5. False negative error rate balance, Equal opportunity	✓
6. Equalized odds	✓
7. Conditional use accuracy equality	✓
8. Overall accuracy equality	✓
9. Treatment equality	✓
10. Test-fairness or calibration	✓
11. Well-calibration	✓
12. Balance for positive class	✓
13. Balance for negative class	✓
14. Causal discrimination (direct discrimination)	*
15. Fairness through unawareness	*
16. Fairness through awareness	✗
17. Counterfactual fairness	✗
18. No unresolved discrimination	✗
19. No proxy discrimination	✗
20. Path based causal reasoning	✗

Figure 1: Bias preservation checklist

Q1: Are you using fairness metrics to solely diagnose disparity, but are not making substantive decisions about individuals?

Yes: Both bias preserving and transforming metrics can be used.
No: Go to Question 2.

Q2: Are you deploying a system to make decisions in an area known to have unacceptable historical social inequality?

Yes: Go to Question 3.
No: Recommend investigation of possible bias in use case before choosing a metric. In cases where historical inequality does not exist, or known disparity has been deemed legally justified, both bias preserving and transforming metrics can be used.

Q3: Are you deploying the system and in a legal jurisdiction that solely promotes formal equality?

Yes: Both bias preserving and transforming metrics can be used.
No: Go to Question 4.

Q4: Are you deploying the system and in a legal jurisdiction that promotes substantive equality?

Yes: Recommend using a bias transforming metric.
No: Both bias preserving and transforming metrics can be used.

Wachter, Mittelstadt, & Russell (2021).

AI / ML researchers are working on the issues. A lot.

(This is the short more algorithm-oriented part of my presentation. Which essentially references Salvatore Ruggieri's excellent work to bring together this complex field in a short presentation.)

So we are back to a general data mining / machine learning setting

- We have a metric
  - Accuracy
  - “Fairness”
    - (once we have chosen one)
- and want to optimise for this.
- or – more generally in this field – trade off two goals (accuracy / fairness)
- How?
  - Pre-processing (modify the data)
  - In-processing (modify the algorithm)
  - Post-processing (modify the results)

Pre-processing

Data sanitization approaches

- Instance reweighting
  - To rebalance distributions
  - F. Kamiran and T. Calders, Data Preprocessing Techniques for Classification without Discrimination. Knowledge and Information Systems, 2012.
- Promotion/demotion of decisions
  - Change decisions to discriminated instances
  - S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans. Knowl. Data Eng., 25(7):1445–1459, 2013.
- Feature distribution distortion
  - F. du Pin Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, K. R. Varshney: Optimized Pre-Processing for Discrimination Prevention. NIPS 2017: 3992-4001
- Generalization
  - Abstract feature values, as for privacy-preserving data disclosure
  - S. Ruggieri. Using t-closeness anonymity to control for non-discrimination. Transactions on Data Privacy. Vol. 7, Issue 2, August 2014, 99-129.

Slide from Salvatore Ruggieri. *Discrimination and Fairness*. NoBIAS OnBoardingWeek, 23 March 2021

In-processing

Fair Classifiers: Algorithms

Calders and Verwer (2010): trains separate Naïve Bayes models for the values and iteratively assesses the fairness of the combined model, makes small changes to the observed probabilities in the direction of reducing the measure, and retrains their two models.

Feldman et al. (2015): a preprocessing approach that modifies each attribute so that the marginal distributions based on the subsets of that attribute with a given sensitive value are all equal.

Kamishima et al. (2012): introduce a fairness-focused regularization term and apply it to a logistic regression classifier.

Zafar et al. (2017): re-express fairness constraints (which can be nonconvex) via a convex relaxation. This allows them to maximize accuracy subject to fairness and also maximize fairness subject to fairness constraints

Zhang et al. (2018): learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions.

Delobelle et al. (2020, 2021). Ethical Adversaries Framework – combines two types of adversarial learning

Slide mostly from Salvatore Ruggieri. *Discrimination and Fairness*. NoBIAS OnBoardingWeek, 23 March 2021

Post-processing

Fairness post-processing

- Involving human experts (*human-in-the-loop*) in the exploration and interpretation of the model or of the model's decisions.
  - Bettina Berendt, Soren Preibusch. Toward Accountable Discrimination-Aware Data Mining:¨ The Importance of Keeping the Human in the Loop - and Under the Looking Glass. In: Big Data 5.2, 2017, pp. 135–152.
- Altering the model's internals, for instance by correcting the confidence of classification rules, or the probabilities of Bayesian models
  - D. Pedreschi, S. Ruggieri, F. Turini. Measuring discrimination in socially-sensitive decision records. SIAM Conference on Data Mining (SDM 2009): 581-592. SIAM, April 2009.
- Altering the model's decision strategy, e.g., by differentiating thresholds for different social groups.
  - Kamiran, F., Mansha, S., Karim, A., & Zhang, X. (2018). Exploiting reject option in classification for social discrimination control. Informa-tion Sciences, 425,18–33
- Prediction-time approaches: correcting their predictions at run-time
  - Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. M. (2018). A reductions approach to fair classification. ICML, 80,60–69.

Slide from Salvatore Ruggieri.
Discrimination and Fairness. NoBIAS OnBoardingWeek, 23 March 2021

Discrimination-aware data mining, fairness-aware data mining, fair machine learning, de-biasing

- Various terms have been and are being used, extremely fast-growing field, many surveys and (fewer) comparative studies, e.g.

Fairness-Aware Machine Learning
An Extensive Overview

Jannik Dunkelau and Michael Leuschel

Heinrich-Heine-Universität Düsseldorf
Universitätsstraße 1 · 40225 Düsseldorf
jannik.dunkelau@hhu.de michael.leuschel@hhu.de

Abstract. We provide an overview of the state-of-the-art in fairness-aware machine learning and examine a wide variety of research articles in the area. We survey different fairness notions, algorithms for pre-, in-, and post-processing of the data and models, and provide an overview of available frameworks.

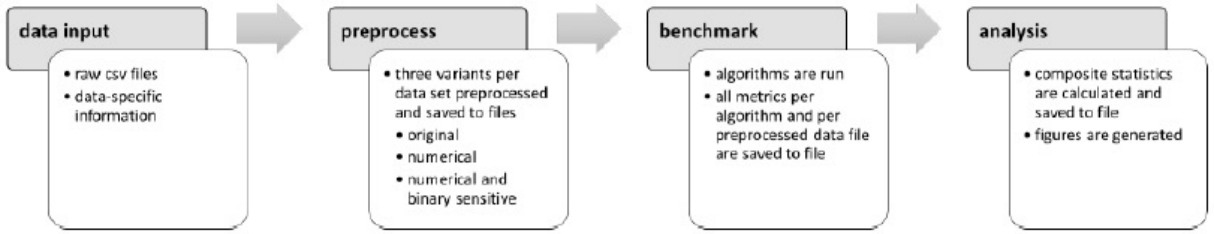
A Survey on Bias and Fairness in Machine Learning

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

[cs.LG] 17 Sep 2019

With the widespread use of AI systems and applications in our everyday lives, it is important to take fairness issues into consideration while designing and engineering these types of systems. Such systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that the decisions do not reflect discriminatory behavior toward certain groups or populations. We have recently seen work in machine learning, natural language processing, and deep learning that addresses such challenges in different subdomain
biases that these appli
different real-world a
biases that can affect A
researchers have defin
different domains and s
in the state-of-the-art r
and solutions that can b
motivate researchers to

Friedler et al. . A comparative study of fairness-enhancing interventions in machine learning. FAT 2019: 329-338. https://arxiv.org/abs/1802.04422



Pyhton code: https://github.com/algofairness/fairness-comparison

Slide based on Salvatore Ruggieri.
Discrimination and Fairness. NoBIAS OnBoardingWeek, 23 March 2021

But is it really a problem of the data + algorithms + result filtering?

- Think back of the Amazon recruiting example.
- And let’s consider more issues.

Feature engineering and data collection: Ex. in the AMS system: “Care obligations”

Care obligations are only asked about for female applicants!

http://epub.oeaw.ac.at/ita/ita-dossiers/ita-dossier052en.pdf

Risk of algorithmic discrimination may arise also from economic considerations

Scharf said the county chose Northpointe's software over other tools because it was easy to use and produced "simple yet effective charts and graphs for judicial review." He said the system costs about $22,000 a year.

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

Discrimination arises from procedures and structures

The case stems from the theft of five watches worth about $4,000 from a Shinola store in the Cass Corridor. A security officer reviewed video footage that showed the suspect wearing a St. Louis Cardinals baseball cap, but the man did not look into the camera, according to the lawsuit.

Michigan State Police ran Williams' photo through facial recognition software, which returned a hit for Williams. Detroit investigators then showed six photographs — including one of Williams' driver's license from six years earlier — to the security officer, who had not witnessed the incident in person, according to a complaint against the police.

"None of us looked alike. And I was like, who put this together?" Williams told lawmakers. "I look nothing like the other guys. I was actually, like, 10 years older and all the other guys in the pictures."

Williams said he doesn't feel like he's ever received a proper explanation for what happened that led to his wrongful arrest.

https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

Discrimination arises from procedures and structures

The case stems from the theft of five watches worth about $4,000 from a Shinola store in the Cass Corridor. A security officer reviewed video footage that showed the suspect wearing a St. Louis Cardinals baseball cap, but the man did not look into the camera, according to the lawsuit.

Michigan State Police ran Williams' photo through facial recognition software, which returned a hit for Williams. Detroit investigators then showed six photographs — including one of Williams' driver's

security officer, who had not

rding to a complaint against the

e, who put this together?"

ng like the other guys. I was actually, like, 10 years older and all the other guys in the pictures."

Williams said he doesn't feel like he's ever received a proper explanation for what happened that led to his wrongful arrest.

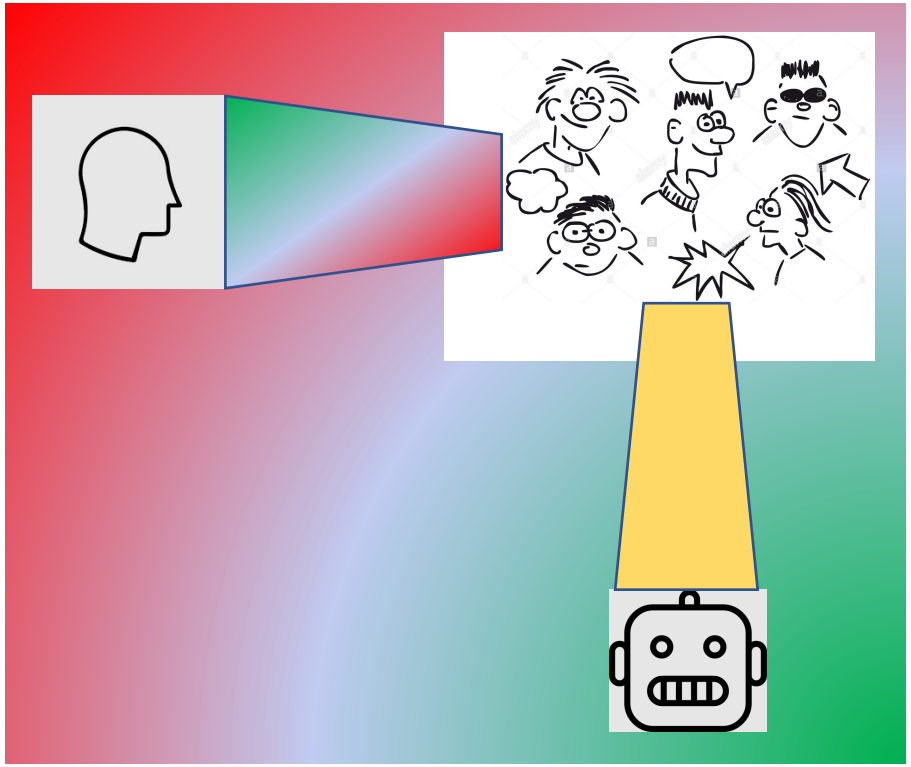
His history with the criminal justice system is what made this incident so scary, he said, because this would have been his third felony, meaning he was at risk of a long sentence. When the prosecutor offered a plea deal, he almost took it even though he was innocent.

https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

Discrimination arises from systems. And: “unfair” or “unjust”?

- Is a (potential) discrimination between social groups really the only / main problem here?
- Systems such as the AMS are based on many assumptions, e.g.
  - Risk factors for unemployment are (nearly) exclusively found in the jobseeker.
  - The categorization of citizens based on their economic potential is politically legitimate and useful.
- The analysis and “repair” of such systems makes further assumptions, such as
  - Risk factors for bias / discrimination are (nearly) exclusively found in the decision-maker (whether human or machine / ADMS).
  - Thus, structural discrimination tends to get neglected.
- AI regulation: Who determines the “risk” of these systems?
- Do we, as a society, want such systems?

  (Scott et al., submitted; Delobelle et al., submitted: AU/BE/PL/DE)



Goals expressed in terms of football

Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity

Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

Justice

All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed.** The systemic barrier has been removed.

Internet meme with original source unknown to me. Can be found for ex. at https://medium.com/the-idiots-guide-to-life/justice-is-not-equality-and-equity-29495532d2ba

No, I have no "solution" for job market questions, but:
AI / ML can also be used very differently!

Why is Swedish the top language in Sweden?

Duolingo's Swedish course turns out to be the most popular in Sweden itself: 27% of all users in Sweden are learning Swedish.

Why would that be the case? Immigration to Sweden has been skyrocketing in recent years: one in six Swedish residents in 2015 was born outside of Sweden. The fastest growing foreign-born groups are from Syria and Afghanistan, reflecting a recent increase in the refugee population. Duolingo recently released a Swedish course for Arabic speakers, which will hopefully help!

UNHCR and Duolingo Aim to Help More Refugees Enroll in Higher Education

Riham Alkousaa / 21 Jun 2021

https://blog.duolingo.com/which-countries-study-which-languages-and-what-can-we-learn-from-it/
https://www.al-fanarmedia.org/2021/06/unhcr-and-duolingo-aim-to-help-more-refugees-enroll-in-higher-education/

Countermeasures: What can AI/ML researchers contribute to an interdisciplinary and collaborative approach that aims at mitigating these issues?

- Formalization of metrics and analysis of their formal properties can enhance clarity about goals.
- Criticism of assumptions in the data and concerning statistical properties → better assess the reliability of predictions
- We know that description is not the same as prediction and can therefore criticize statements such as “this only reflects the harsh realities of the labour market” as oversimplifications.
- Criticism of interpretation: Data about the effectiveness/efficiency of AMS training measures, i.e. the intervention, are missing (and wouldn’t that be what one should optimize for?)
- Systems thinking!

So what should hide, and what should we expose, to be fair?

You can start today – by engaging with the current proposal for AI regulation in the EU !

Task:
Imagine you are a malicious AI developer. How would you work your way around this regulation (assuming it became law in ints present formulation)?

Proposal for an ARTIFICIAL INTELLIGENCE ACT, some excerpts

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

- **Title II** establishes a list of prohibited AI. The regulation follows a risk-based approach, differentiating between uses of AI that create (i) an unacceptable risk, (ii) a high risk, and (iii) low or minimal risk. The list of prohibited practices in Title II comprises all those AI systems whose use is considered unacceptable as contravening Union values, for instance by violating fundamental rights.
- the classification as high-risk does not only depend on the function performed by the AI system, but also on the specific purpose and modalities for which that system is used.
- (Frequent references to the risks of discrimination and the need to protect against this.)

Article 3 Definitions

- For the purpose of this Regulation, the following definitions apply:
- (1) 'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with;
- (18) 'performance of an AI system' means the ability of an AI system to achieve its intended purpose;
- *Article 15 Accuracy, robustness and cybersecurity*
- 3. High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way to ensure that possibly biased outputs due to outputs used as an input for future operations ('feedback loops') are duly addressed with appropriate mitigation measures.

Q & A at ACDL 2021, postprocessed & with references

- Q: How do you define white or black (what if the person is mixed)?
  - A: In general, that's indeed one of the key problems of the setup per se: How are the categories defined, and who defines them? See the excellent analysis by Barocas & Selbst (2016) of the importance of such basic choices (before we even go to algorithms)
  - Specifically in the example of redlining: The US has employed definitions such as "One Drop of (Black) Blood makes you black" (see for example https://us.macmillan.com/books/9780374527945) and self-attribution (in modern census questions).
  - Similar questions arise when we go from binary gender to non-binary gender.
- Q: This discrimination also ocurrs to rich/poor people. Should we also hide this data?
  - A: Indeed, socio-economic status is an important factor in discrimination (see e.g. Eubanks, 2018), although it could be argued that current societal consensus often does not regard this as quite as objectionable as for example differentiation by gender or so-called 'race' (see for example Binns, 2018, for a further discussion, and also the implicit idea that the unemployed are somehow "responsible" for their lot, cf. slide 50). Whether we should hide such data or not … can we (if so much else is correlated with it?)
- Q: The distinction between "what is fair" and "what is legal" is not clear.
  - A: Indeed, our "ethical" and "moral" notions of what is fair and our legal notions of what is fair, while overlapping, are not identical. And the philosophical, societal, and legal "operationalisations" also change over time, and of course are different across and even within societies. In discrimination-aware data mining, fairness-aware machine learning, etc., we generally have to assume we know the factors (see Berendt & Preibusch, 2014, and Berendt, in press, for further discussion).
  - Theoretically we could look at any grounds or combination of them as a source of unfairness (Kearns et al., 2018). In the limit, if we do not want to make any differentiation, any decision or choice is unfair (why even do machine learning then? But … there is so much discrimination by the "classical" grounds: if we even manage to improve things here, we'll have made a HUGE contribution.
- Q: Currently, I'm doing some research on how to increase User Experience. I thought about doing an experiment With male/female participants to get some differentiation. Am I being sexist? According to this lecture, should I ignore this differentiation and let the data separate the results as male/female by itself (Indirect discrimination)?
  - A: It is not the case that any differentiation by a protected attributed is discrimination. On the contrary, it may be necessary to differentiate in order to NOT discriminate (cf. Berendt & Preibusch, 2014). Also, investigating possible differences is not necessarily treating people differently; on the contrary, in many human-subjects experiments, you want to have balancing across gender. The question is what your goals are. What do you mean by "getting some differentiation"? And why are you so sure that you will obtain results that constitute different treatment that disproportionally disadvantages one gender. So, in sum, I'd need to know more detail to answer this question. Email me ☺
- Q: is this [the data privacy metrics such as k-anonymity, l-diversity and to-closeness[related to differential privacy?
  - A: Yes. See https://www.dbs.ifi.lmu.de/Lehre/KDD/SS16/skript/8_PrivacyPreservingDataMining.pdf for an intro overview and this paper for relevant further insights:

  Josep Domingo-Ferrer, David Sánchez, Alberto Blanco-Justicia: The limits of differential privacy (and its misuse in data release and machine learning). Commun. ACM 64(7): 33-35 (2021). Available at https://arxiv.org/abs/2011.02352

Literature (1)

Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. Der AMS Algorithmus - Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Technical Report. Wien: ÖAW, ITA 2020-02. http://epub.oeaw.ac.at/?arp=0x003bfdf3/

Earlier version of the work in English: Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, Astrid Mager: Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. Frontiers Big Data 3: 5 (2020) https://www.frontiersin.org/articles/10.3389/fdata.2020.00005/full

Solon Barocas, Moritz Hardt and Arvind Narayanan (2019). Fairness and Machine Learning. http://www.fairmlbook.org

Barocas, Solon and Selbst, Andrew D., Big Data's Disparate Impact (2016). 104 California Law Review 671 (2016), Available at SSRN: https://ssrn.com/abstract=2477899 or http://dx.doi.org/10.2139/ssrn.2477899

Berendt, B. (in press). Algorithmic discrimination. To appear in G. Comandè (Ed.), *Encyclopedia of Law and Data Science.* Edward Elgar Publishing Ltd. (last author version: https://people.cs.kuleuven.be/~bettina.berendt/Papers/Algorithmic Discrimination BB 2020 10 13.pdf)

Bettina Berendt, Sören Preibusch: Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. Artif. Intell. Law 22(2): 175-209 (2014). https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_preibusch_2014.pdf

Berendt, B. & Preibusch, S. (2017). Toward accountable discrimination-aware data mining: The importance of keeping the human in the loop – and under the looking-glass. Big Data, 5(2), 135-152. https://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_preibusch_2017_last_author_version.pdf

Binns (2018). Fairness in Machine Learning: Lessons from Political Philosophy. FAT* / Proceedings of Machine Learning Research 81:1. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3086546

Joy Buolamwini, Timnit Gebru: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT 2018: 77-91. https://www.semanticscholar.org/paper/Gender-Shades%3A-Intersectional-Accuracy-Disparities-Buolamwini-Gebru/18858cc936947fc96b5c06bbe3c6c2faa5614540

Citation on p.8 from https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

Simon Caton, Christian Haas: Fairness in Machine Learning: A Survey. CoRR abs/2010.04053 (2020). https://arxiv.org/abs/2010.04053

Custers, B., Calders, T., Schermer, B., Zarsky, T. (Eds.) (2013). Discrimination and Privacy in the Information Society. Data Mining and Profiling in Large Databases. Springer.

Literature (2)

Pieter Delobelle, Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, David Hartmann, Tianling Yang, Elena Murasso, Karolina Sztandar-Sztanderska and Bettina Berendt (submitted). Time to Question if We Should: Data-Driven and Algorithmic Tools in Public Employment Services. *Available from the authors per email request.*

Pieter Delobelle, Paul Temple, Gilles Perrouin, Benoît Frénay, Patrick Heymans, Bettina Berendt: Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning. SIGKDD Explor. 23(1): 32-41 (2021). Earlier version 2020 available at https://arxiv.org/abs/2005.06852

V Eubanks, Automating Inequality. How High-Tech Tools Profile, Police and Punish the Poor (St. Martin's Press 2018)

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Transactions on Information Systems 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

Ben Hutchinson, Margaret Mitchell: 50 Years of Test (Un)fairness: Lessons for Machine Learning. FAT 2019: 49-58. https://arxiv.org/abs/1811.10104

Michael J. Kearns, Seth Neel, Aaron Roth, Zhiwei Steven Wu: Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. ICML 2018: 2569-2577. http://proceedings.mlr.press/v80/kearns18a.html

Salvatore Ruggieri (2014). Using t-closeness anonymity to control for non-discrimination. Transactions on Data Privacy, 7, 99-129. https://www.semanticscholar.org/paper/Using-t-closeness-anonymity-to-control-for-Ruggieri/58786467d9e5085d630f1575ebcf444fe4f0c36b

Kristen M. Scott, Pieter Delobelle, Sonja Mei Wang, Milagros Miceli, David Hartmann, Tianling Yang, Elena Murasso, Karolina Sztandar-Sztanderska and Bettina Berendt (submitted). Data-Driven and Algorithmic Tools in Public Employment Services: Towards a Stakeholder-Centric Perspective. *Available from the authors per email request.*

Wachter, Mittelstadt, & Russell (2021). Bias Preservation in Machine Learning: The legality of fairness metrics under EU non-discrimination law. West Virginia Law Review, forthcoming. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772

I. Zliobaite , B. Custers . Using sensitive personal data may be necessary for avoiding discrimination in data driven decision models. Artif . Intell . Law 24(2): 183 201, 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3047233

Plus the references on pp. 41ff.!