Fairness

Magnus Nielsen, SODAS, UCPH

Agenda

- What is fairness?
- Algorithmic decisions
- Auditing
- Statistical fairness measures
  - Independence
  - Separation
  - Sufficiency
- Achieving fairness

Introduction

This session will be based mainly on the book Fairness and Machine Learning: Limitations and Opportunities (Barocas et al., 2019)

- Available online at fairmlbook.org
- Good starting point for discussing fairness
- Updated in 2022

A survey on more aspects and applications can be found in Mehrabi et al. (2021)

What is fairness?

Illuminate we will!

Ursula Franklin (Olteanu et al., 2019):

- *“For your own sanity, you have to remember that not all problems can be solved. Not all problems can be solved, but all problems can be illuminated”*

Question

In your opinion, what is fairness in the context of machine learning? What is bias?

- Does it depend on the situation? Algorithm?

Fairness is a societal concept

Chouldechova (2017):

- *“It is important to bear in mind that fairness itself (…) is a social and ethical concept, not a statistical concept.”*

Nevertheless, we’ll try to turn it into observable statistical measures towards the end

But first, bias

Bias is a broad term with many different interpretations

- Often more granular and related to a single phenomena
- Lists of different biases can be seen in Mehrabi et al. (2021) and Olteanu et al. (2017)

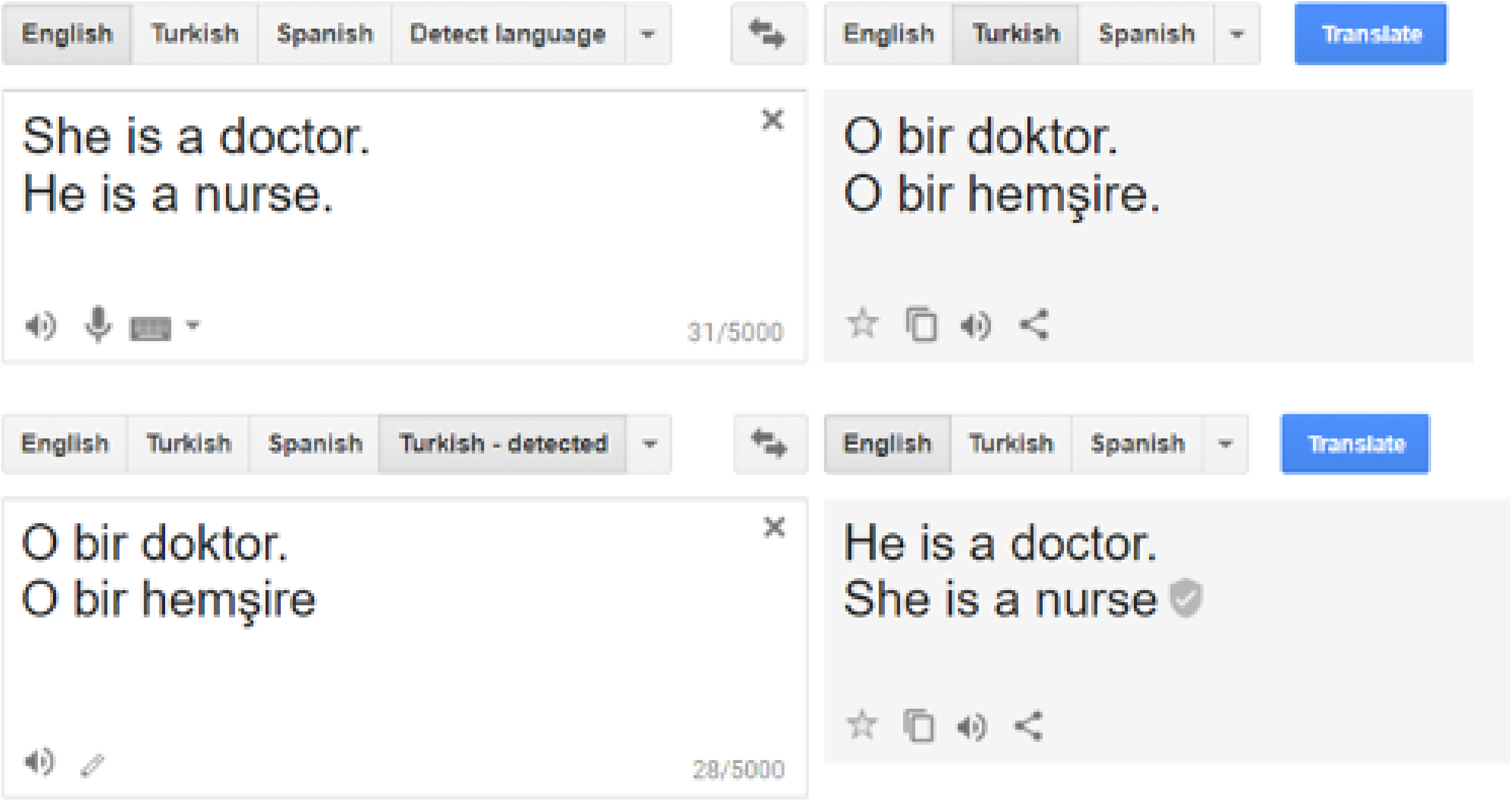
Some people try to avoid this term due to the ambiguity, including Barocas et al. (2019)

Some examples

The list is long

- Measurement bias
- Omitted variable bias
- Sampling bias
- Aggregation bias
- Self-selection bias

Unfair, biased or both?



She is a doctor.
He is a nurse.

O bir doktor.
O bir hemşire.

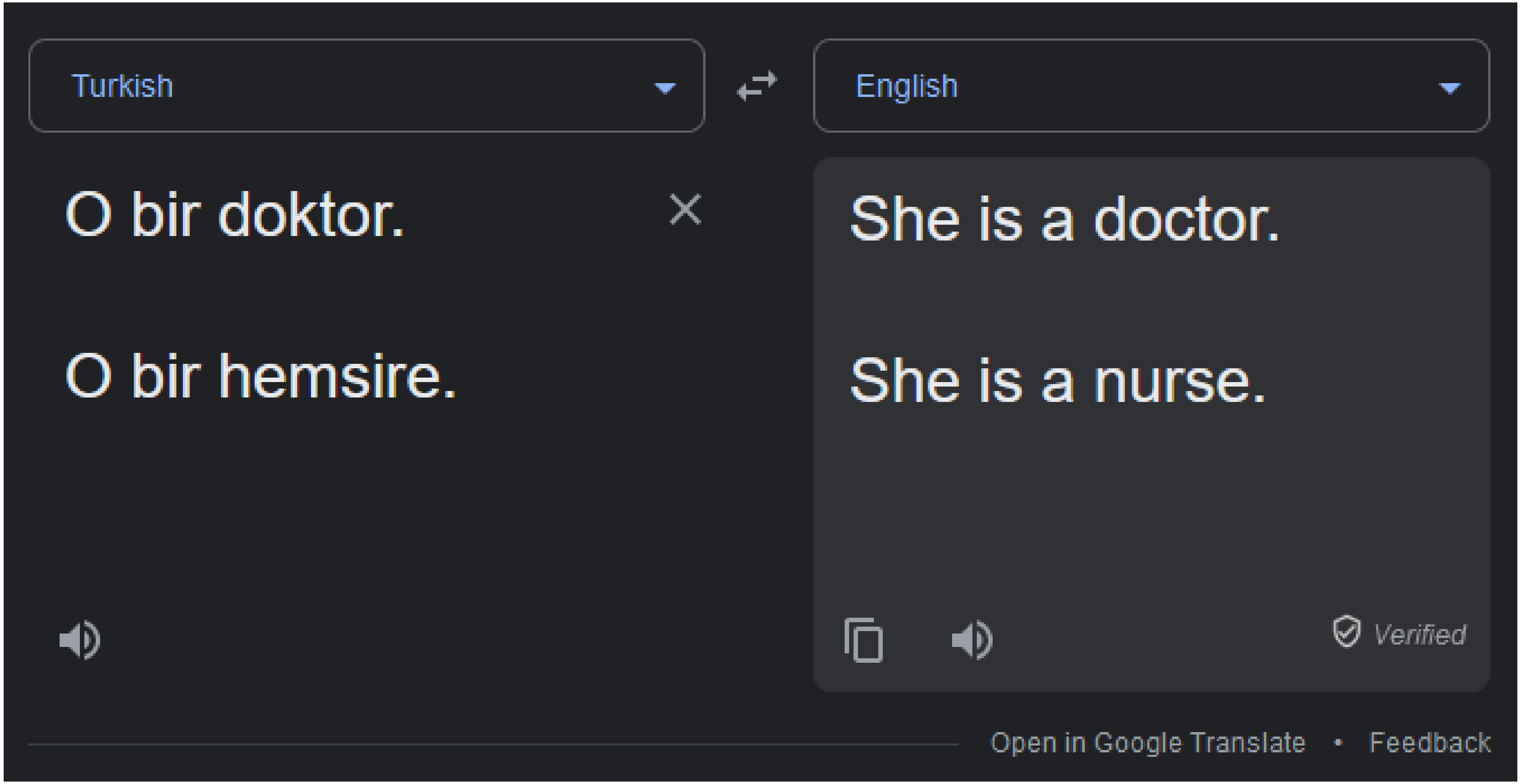
O bir doktor.
O bir hemşire

He is a doctor.
She is a nurse

Gender stereotypes

Source: Barocas et al. (2019)

Problem… solved?



O bir doktor.

O bir hemsire.

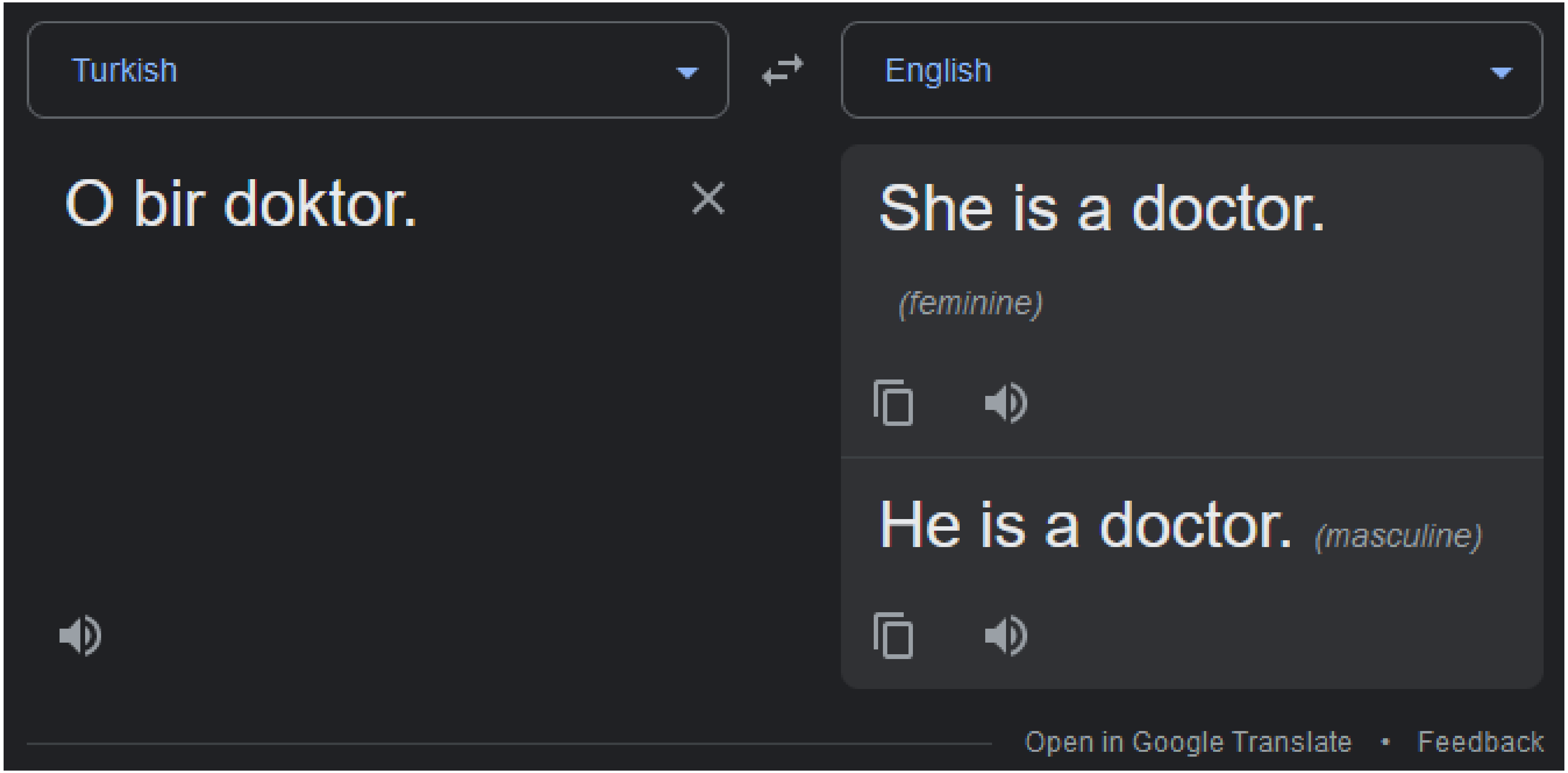
She is a doctor.

She is a nurse.

Default gender

Source: Google Translate

Problem solved!



Turkish

English

O bir doktor.

She is a doctor. (feminine)

He is a doctor. (masculine)

Multiple options

Source: Google Translate

Algorithmic fairness

Algorithmic fairness is a relatively new concept

- Fairness as a concept has existed for a long time
- Algorithms have existed for a long time

Lots of information on fairness, but generally not with respect to algorithms

Fairness & decision-making

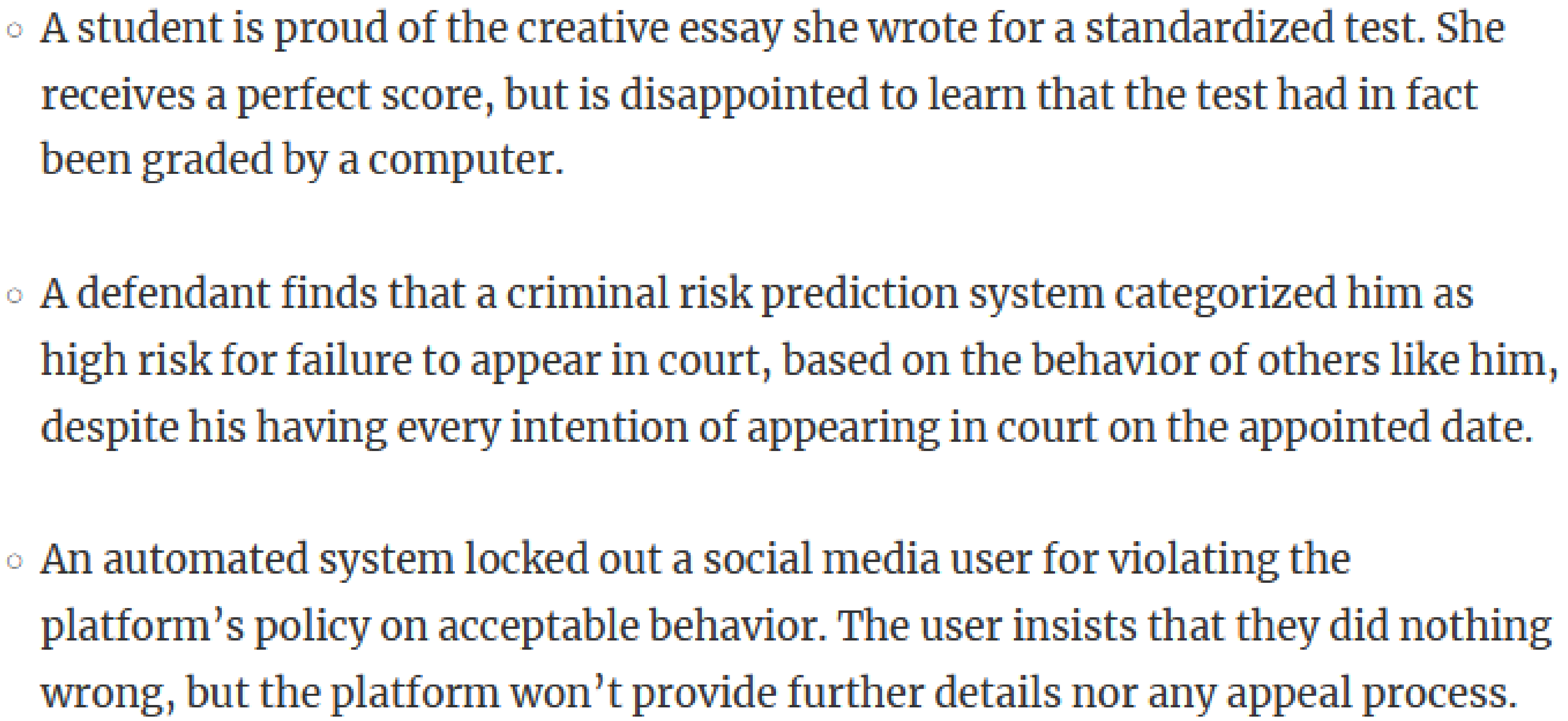
An *'absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics'* (Mehrabi et al., 2021)

We will consider group fairness

- No discrimination based on sensitive attributes such as race, gender, ethnicity

For a more thorough walkthrough of moral notions of fairness, see chapter 4 in Barocas et al., 2019

Unfairness?



- A student is proud of the creative essay she wrote for a standardized test. She receives a perfect score, but is disappointed to learn that the test had in fact been graded by a computer.

- A defendant finds that a criminal risk prediction system categorized him as high risk for failure to appear in court, based on the behavior of others like him, despite his having every intention of appearing in court on the appointed date.

- An automated system locked out a social media user for violating the platform’s policy on acceptable behavior. The user insists that they did nothing wrong, but the platform won’t provide further details nor any appeal process.

Example situations

Source: Barocas et al. (2019)

Legitimacy

Is it right to deploy a machine learning algorithm in the first place?

- This we call legitimacy

Precedes other fairness concerns

- No algorithm $\rightarrow$ no algorithmic fairness concerns

Algorithmic decisions

What are algorithms replacing?

Humans are subject to subjectivity, arbitrariness, and inconsistency

- Historically replaced by bureaucracies

Algorithms most commonly replace bureaucracies

- Seldom are important decisions in settings of social-scientific importance made by a single human

Legitimacy

Public institutions are in general more regulated than private companies

- Higher need for legitimacy

To increase legitimacy, consider things such as

- Transparency
- Relevance of inputs
- Possibility for recourse

Three situations

Automating pre-existing decision-making rules

- Hardcoding, not machine learning

Learning decision-making rules from data on past decisions in order to automate them

- Target is given by previous decisions

Deriving decision-making rules by learning to predict a target

- New target to be decided upon

Question

You’re tasked with creating a new centralized admissions system for higher education using machine learning.

- What is your target of choice?
- Does it depend on your employer?

Fixed decision-makers

Algorithmic decisions as a process of inductive reasoning

- Choice of target important

Some issues to consider:

- Overfitting
  - We have tools to combat this
- Distribution shifts
  - Hot topic in research

Auditing

Examining decision-makers

Given a

- notion of fairness
- decision-maker (black box)

We can audit the decision-maker!

Classic auditing

Ayres & Siegelman (1995)

- Send ‘identical’ testers to bargain for cars
- Vary race
- Black males receive final offers that are $1.100 more than white males

Bertrand & Mullainathan (2004)

- Send identical job applications to apply for jobs
- Vary name
- Traditionally white names 50% more likely to receive callbacks

Question

What could be reasons that firms engage in this sort of discrimination?

- Is it fair?

Fairness as blindness

A wish for null findings implicate no discrimination based on sensitive attribute

- Based on both observables (somewhat) and unobservables
- This subsumes a definition of fairness as blindness

Possible to illuminate decision making with partial dependence functions

Statistical fairness measures

Binary classification

We have

- Observed set of target and features, $\{Y, X\}$
- X contains a sensitive attribute A
- A risk score generated by a classifier

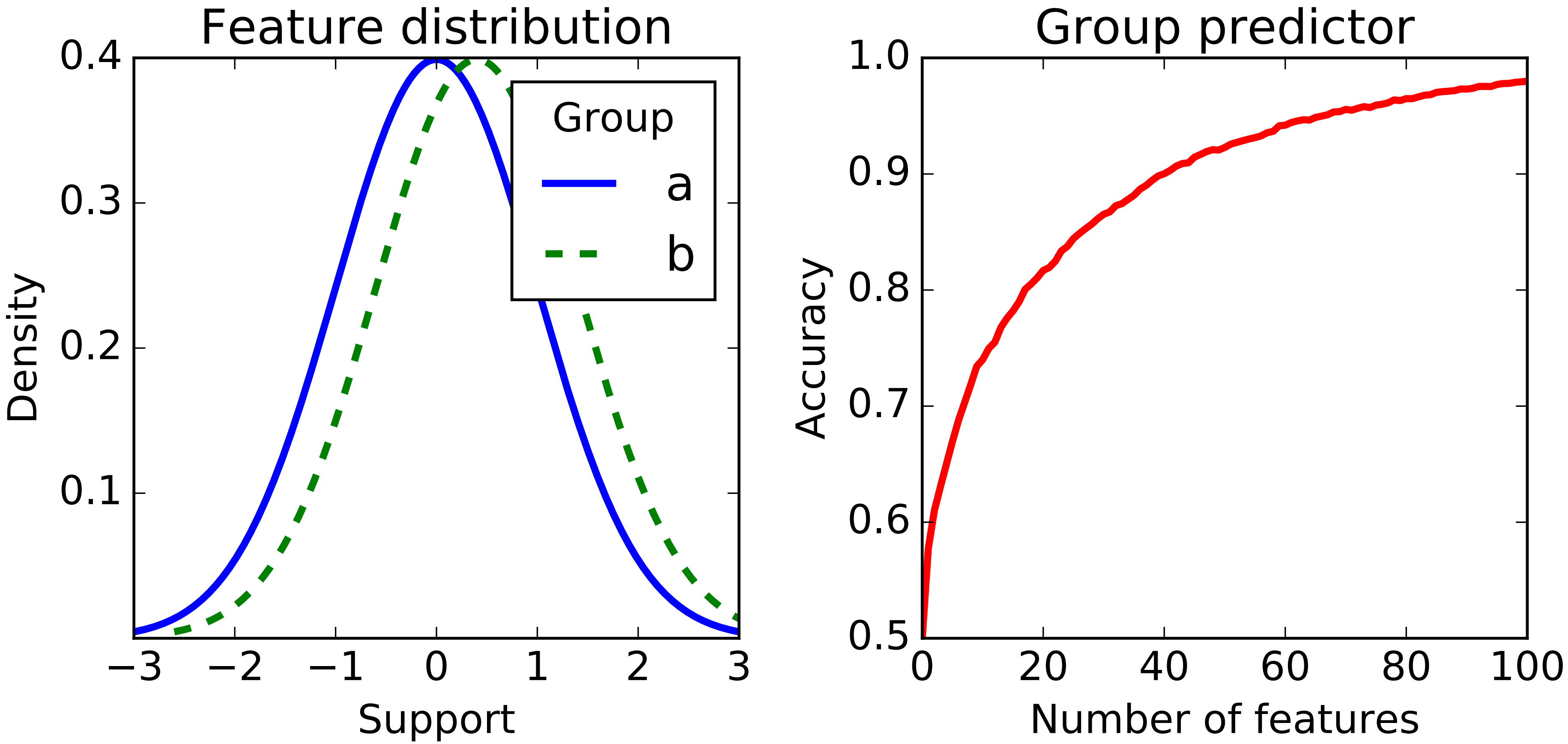
$$r(x) = P\{Y = 1 \mid X = x\}$$

Based on this risk score, we can calculate a predicted outcome Y

Question

Can we obtain fairness by withholding information about sensitive attributes?

No fairness through unawareness



Multiple correlated variables captures bias

Source: Barocas et al. (2019)

Different types of fairness

Fundamentally three types of fairness metrics

- Acceptance rates, $P\{\widehat{Y} = 1\}$
- Error rates, $P\{\widehat{Y} = 0 \mid Y = 1\}$ and $P\{\widehat{Y} = 1 \mid Y = 0\}$
- Outcome frequency given risk score, $P\{Y = 1 \mid R = r\}$

These are equalized across groups defined by the sensitive attribute

Loan applications

We will consider the same example for our fairness metrics

Given

- Information on gender (A), income (and other things) (X) and outcome (Y)

Decision to make: Who should get a loan?

Question: What target should we use to create our model?

Independence

Acceptance rate parity

‘True’ equality

- Sensitive attribute is unconditionally independent of score
- $A \perp R$
- Equal acceptance across groups

This is also commonly referred to as demographic parity

Rewriting

Easy to work with algorithmically

$$P\{\widehat{Y}|A = a\} = P\{\widehat{Y}|A = b\}$$

In practice:

- Calculate mean of $\widehat{Y}$ across sensitive attribute values
- Assert if they differ

Question

Could independence have adverse consequences for either group in the loans context?

- Consider the qualifications of the applicants
- How could one achieve independence?

Relaxations

$$P\{\widehat{Y}|A=a\} \geq P\{\widehat{Y}|A=b\} - \epsilon$$

$$\frac{P\{\widehat{Y}|A=a\}}{P\{\widehat{Y}|A=b\}} \geq 1 - \epsilon$$

The U.S. has an 80% rule to detect discriminatory hiring

- Protected groups hired at least 80% as much as white men
- Feldman et al. (2015) argue that $\epsilon = 0.2$ encapsulates this
- *'Supreme Court has resisted a "rigid mathematical formula"'* (Feldman et al., 2015)

Some considerations

No influence if groups differ in covariates and outcomes

- Normative question if desired

Can lead to adverse outcomes, e.g.

- Increase in acceptance for minority group
- Can create more false positives
- Creates bad track record

Note that minority groups by definition have less training data

Question

Assume that men overall are worse at paying back loans and defaulting is costly

- Just assumptions
- No assumptions as to why

Could this justify some discrimination based on gender?

Separation

Error rate parity

Equality within error rates

- Score independent of sensitive attribute given outcome
- $R \perp A|Y$
- Equalization of errors made within each strata defined by true outcome

Also known as equalized odds

Rewriting for binary classifier

$$P\{\widehat{Y} = 1 | Y = 1, A = a\} = P\{\widehat{Y} = 1 | Y = 1, A = b\}$$

$$P\{\widehat{Y} = 1 | Y = 0, A = a\} = P\{\widehat{Y} = 1 | Y = 0, A = b\}$$

Equality of

- False negative rate
  - Thereby also true positive rate
- False positive rate
  - Thereby also true negative rate

Why?

Generally misclassification has a cost, e.g. a lost opportunity

- Higher error rates amongst disadvantaged groups cause further harm

Note that target variables can encode previous inequality and injustice

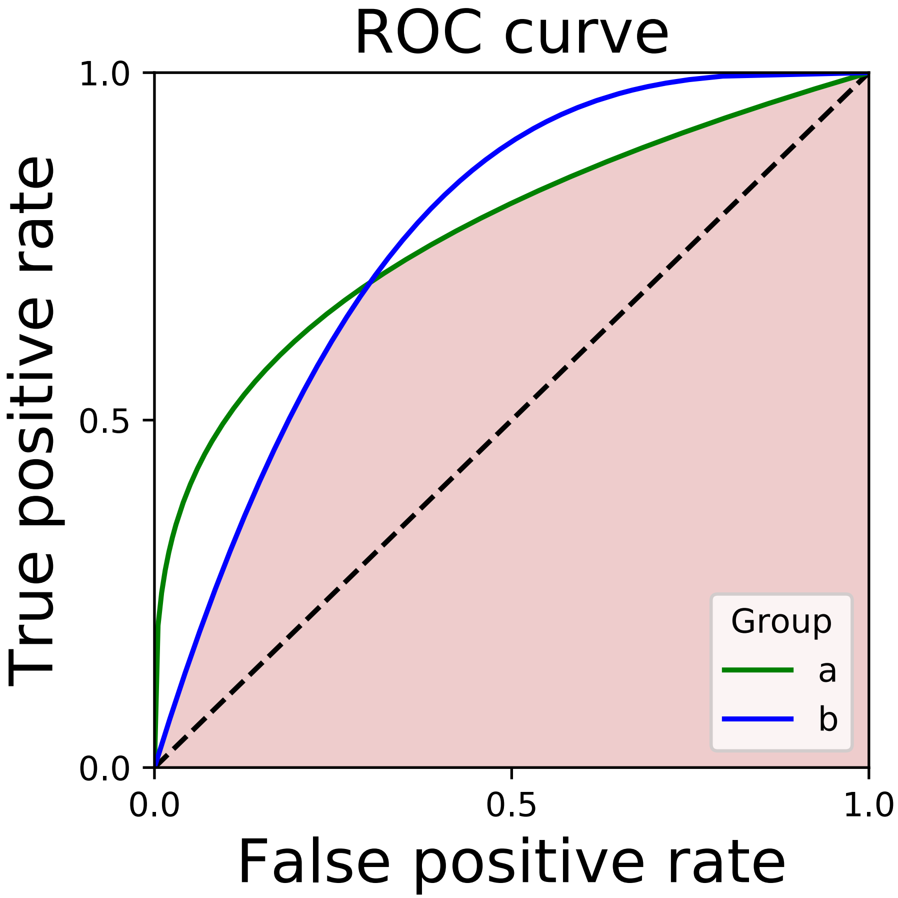
- Separation doesn’t create fairness, but this is more general to supervised learning as a whole

Question

What are the implications of separation for our loan model?

- Is this fair?

Visually



Achievable false and true positive rates

Source: Barocas et al. (2019)

Achieving separation

Tools to achieve red area in figure achievable by a combination of

- Linear combinations of the classifier and classifiers that always predict either true or false
- Different thresholds

Relaxations

Equality of either false negative rate or false positive rate

Consider the costs of misclassification and who experiences them

An example: Screening for job interviews

- Cost to applicant if false negative (denied opportunity)
- Cost to firm if false positive (wasted time)
- Perhaps equality of false negative rates most important

Question

Do you believe a relaxation is appropriate in the loan example?

Sufficiency

Calibrated risk scores

Given a risk score, groups should not differ in outcomes

- Outcome independent of sensitive attribute given risk score
- $Y \perp A | R$

If a model predicts a high risk, then it should be the same high risk for both groups

- No additional information in the sensitive variable

Rewriting

$$P\{Y = 1 \mid R = r, A = a\} = P\{Y = 1 \mid R = r, A = b\}$$

A flexible model which uses A as an input generally (approximately) achieves this

- If the sensitive attribute has predictive power, the model utilizes this information

Fairness through awareness

- In contrast to fairness through unawareness

Question

What are the implications of sufficiency for our loan model?

- Is this fair?

Incompatibilities

A choice to make

We are generally not able to satisfy all fairness criteria at once!

- Trade-offs
- We must consider what is appropriate in our case

The proofs for the following three incompatibilities are found in Barocas et al. (2019)

Independence and sufficiency

Assume $A \not\perp Y$

- Different rates of positive outcomes

Then independence and sufficiency cannot both hold

- Sufficiency requires that the additional information from A encoded in risk score, but independence requires that score and sensitive attribute are uncorrelated

Independence and separation

Assume Y binary, $A \not\perp Y$ and $R \not\perp Y$

- Different rates of positive outcomes and risk score has predictive power

Then independence and separation cannot both hold

- Independence requires that risk score independent of sensitive attribute, but this causes unequal error rates when baselines differ

Separation and sufficiency

Assume $A \not\perp Y$ and all events in the joint distribution (R, A, Y) has positive density

- Different rates of positive outcomes and risk score never fully resolves uncertainty

Then separation and sufficiency cannot both hold

- Sufficiency requires that information about sensitive attribute encoded in score, but then different error rates occur

Achieving fairness

Timing

When in the process to achieve fairness

- Pre-processing, e.g. removing information about sensitive attributes
- In-processing, e.g. constraints based on fairness
- Post-processing, e.g. combining with other classifiers

Doing only fairness assessment and no mitigation is common

Post-processing

Post-processing is common

- Relies explicitly on group membership
- Will be examined in exercises

Comes with some (near) optimality guarantees

Observational criteria

We’ve now looked at different criteria of fairness based on observables

- Many more exist, see table in bottom of ‘Classification’ in Barocas et al. (2019)

Note that this doesn’t tell us anything about mechanisms or causes

- Is it the decision-maker?
- Society as a whole?

How to learn about mechanisms

Attempts with causality

- SCM’s: Causal discovery is not easy in the social sciences
- PO: You need a research design

A reminder

Once again remember Chouldechova (2017):

- *“It is important to bear in mind that fairness itself (…) is a social and ethical concept, not a statistical concept.”*

References

Ayres, I., & Siegelman, P. (1995). Race and gender discrimination in bargaining for a new car. The American Economic Review, 304-321.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American economic review, 94(4), 991-1013.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 259-268).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2, 13.

To the exercises!