Fairness and Biases in ML

Franziska Boenisch and Adam Dziedzic
Course on Trustworthy Machine Learning

Today we turn to fairness in ML – this becomes a crucial topic as ML is used in more high stake applications, such as hiring decisions and deciding whether someone should get a loan.
There are clear benefits to algorithmic decision-making; unlike people, machines do not become tired or bored and can take into account orders of magnitude more factors than people can. However, like people, algorithms are vulnerable to biases that render their decisions “unfair”.

Why does ML need to be fair?

NEWS & COMMENTARY

Why Amazon's Automated Hiring Tool Discriminate

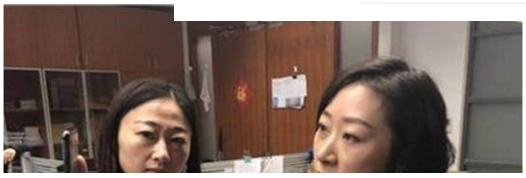
3. Google Translate translated English, gender-neutral, sentences to Turkish phrases that carried heavy gender stereotypes, such as "the doctor" being translated to the male form of the noun (Caliskan et al., 2017), as illustrated in Fig. 1.

Chinese woman offe allows colleague to

Listen to this article

Zhuang Pinghui in Beijing + FOLLOW

Published: 4:00pm, 14 Dec 2017



ı, inferior

his aspect is agnose skin can- s do not include r results when

n-white patients.

2

Examples of unfair results by ML:

> In 2014, a team of engineers at Amazon began working on a project to automate hiring at their company. Their task was to build an algorithm that could review resumes and determine which applicants Amazon should bring on board. But, according to a Reurters report this week, the project was canned just a year later, when it became clear that the tool systematically discriminated against women applying for technical jobs, such as software engineer positions.

How did it do that:

- Downvote candidates who went to certain women's colleges
- Penalize applications that contain the word woman
- Etc.

Why can that happen?

> Models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

https://www.reuters.com/article/idUSKCN1MK0AG/

Or vision models trained to detect skin cancer. They are trained on white skin. Their detection rate becomes poor when applied to black skin.

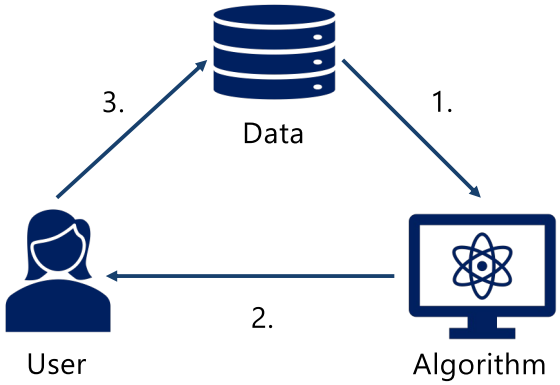
Training datasets in general contain mainly white/light skin, often no pictures from people with African, Afro-Caribbean, or South Asian backgrounds. Hence, healthcare for these groups becomes worse.
https://healthcare-in-europe.com/en/news/ai-in-skin-cancer-detection-darker-skin-inferior-results.html

Similar discrimination can be observed also in the proprietary sector, such as with FaceID in the iPhone.
This has shown to have inferior performance on Asian pheno types, sometimes up to the degree of identifying one person as another one.

Turkish translation:
Turkish is a gender neutral language. However, when we translate it back to English, it has the biases.

How can all this happen? This is because the ML systems:
1. Reproduce the biases they already see from their training data (Amazon hired already more men than women)
2. Are fed data without taking fairness into account (their training data selection is biased)
3. Are tested on only subsets of the distribution that represents the world, but released if they do well on these subsets.

What types of biases do exist?

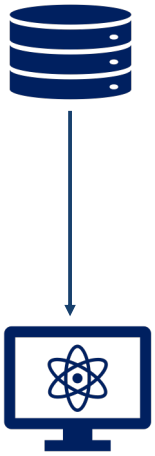


3

Biases can enter ML systems in different ways, connected to the users, the data they produce or the data they interact with, and the influence of that data to the algorithm.
We are going to see the three ways more in detail.

Biases I: data to algorithm

1. **Measurement (reporting) bias**: through choice, utilization, and measurement of particular features.
2. **Omitted variable bias**: ignorance of important outside variables that influence the outcome of the algorithm.
3. **Representation bias:** incorrect sampling of data from a population.
4. **Aggregation bias:** false conclusions about individuals from observing entire population.



4

1. **Measurement or reporting bias:** arises from how we choose, utilize, and measure particular features. — this is related to the amazon hiring example which extracted certain features, such as being a women, or writing the word woman as bad.
2. **Omitted variable bias** occurs when one or more important variables are left out of the model: e.g. a model to predict whether people will cancel their subscriptions for a service. Imagine now, at some point a new competitor arises on the market and offers the same service for half the price. Then, people will quit, and the model will not know because it was not prepared for that
3. **Representation bias** arises from how we sample from a population during data collection process: we sample incorrectly. For example, in the skin cancer detection example, people have just not sampled people with darker skin for training and evaluation.
4. **Aggregation bias** (or ecological fallacy) arises when false conclusions are drawn about individuals from observing the entire population: e.g. seen in clinical aid tools: to detect diabetes, one measures the HbA1c levels, however, these differ in complex ways across genders and

ethnicities, hence, just taking an average will not help the population (even if we have a perfectly balanced dataset)

Biases II: algorithm to user

1. **Algorithmic bias**: through choice of optimization function, regularization, choice of population.
2. **User interaction bias**: e.g. presentation bias (only presented items can be chosen), or ranking bias (top ranked results will become even more popular).
3. **Popularity bias:** popular items get interacted with more, but can be popular through manipulation (bots, fake reviews,...).
4. **Emergent bias:** population changes.
5. **Evaluation bias**: use of inappropriate and disproportionate benchmarks.

5

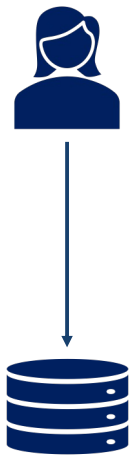
1. Algorithmic Bias. Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm, such as through optimization functions, choices of regularization, and on what part of the population to do the analysis
2. User Interaction Bias.
    1. Presentation Bias. Presentation bias is a result of how information is presented [9]. For example, on the Web users can only click on content that they see, so the seen content gets clicks, while everything else gets no click. And it could be the case that the user does not see all the information on the Web
    2. Ranking Bias. The idea that top-ranked results are the most relevant and important will result in attraction of more clicks than others. This bias affects search engines [9] and crowdsourcing applications
3. Popularity Bias. Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation—for example, by fake reviews or social bots
4. Emergent Bias. Emergent bias occurs as a result of use and interaction with real users. This bias arises as a result of change in population,

cultural values, or societal knowledge usually some time after the completion of design

5. Evaluation Bias. Evaluation bias happens during model evaluation. This includes the use of inappropriate and disproportionate benchmarks for evaluation of applications

Biases III: user to data

1. **Historical bias**: reproduce the unfairness from the past.
2. **Population bias**: user population is different from target population.
3. **Self-selection bias:** a subtype of the selection or sampling bias in which subjects of the research select themselves.
4. **Social bias:** the judgement of other(s) affects our judgement.



6

1. Historical Bias. Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection — this is in the Amazon hiring example: they hired more men in the past, they had certain patterns which are then preferred by the algorithm
2. Population Bias. Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population of the platform from the original target population. One could argue that this underlies the problem of the iPhone in China. It was designed originally for the US market.
3. Self-Selection Bias. Self-selection bias is a subtype of the selection or sampling bias in which subjects of the research select themselves. An example of this type of bias can be observed in an opinion poll to measure enthusiasm for a political candidate, where the most enthusiastic supporters are more likely to complete the poll.
4. Social Bias. Social bias happens when others' actions affect our judgment. An example of this type of bias can be a case where we want to rate or review an item with a low score, but when influenced by other high ratings, we change our scoring thinking that perhaps we are being

too harsh. This happens also in reviewing papers in academia

Different ways to measure fairness

Individual

Group

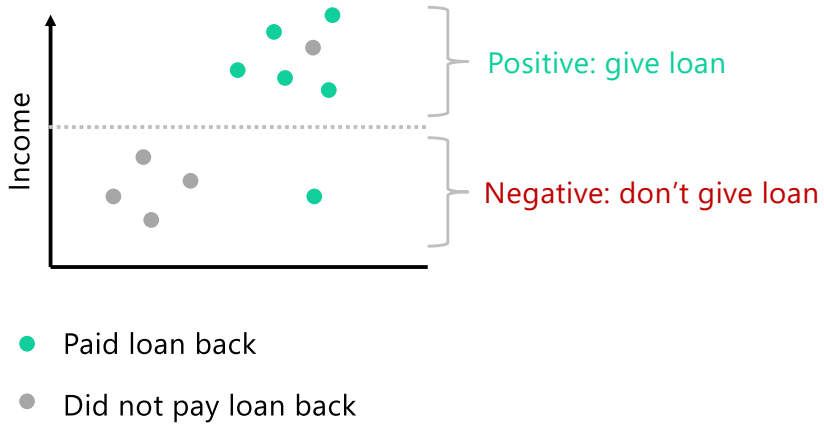
Examples:
- Fairness through awareness
- Fairness through unawareness

Examples:
- Demographic parity
- Equalized odds
- Equal opportunity
- Treatment equality

7

Some fairness definitions center around one individual and making their treatment fair, others focus on entire groups (or subgroups).
All these definitions are currently out there, and there is no right or wrong definition, but some definitions are more suited for some application than others for given applications.
Also note that it is impossible to satisfy some of the fairness constraints at once except in highly constrained special cases.

Running example: Who pays their loan?



- Paid loan back
- Did not pay loan back

8

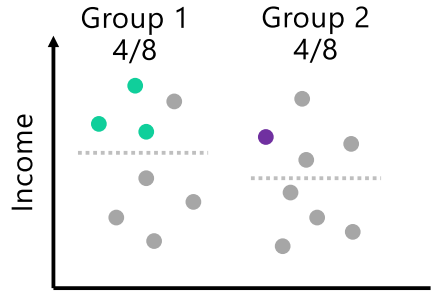
We're going to consider a very simple "ML" model that could be deployed by a bank. It takes as an income the salary of a person, and outputs whether the person gets a load or not. It is based on historic data.

We're going to denote by the full color cycle that someone paid back their loan, whereas the grey cycles indicate that the person did not pay back.

Based on this historic data, we now want to find an income threshold that the bank should consider in the future when deciding who gets a loan. You can see that if we draw a line just in the middle, we might get some individuals wrong (i.e. don't give the loan to someone who deserves it, but give it to someone who won't be able to pay it back.) But overall, that would, based on our historic data be the "best" approach for the bank.

Demographic parity (group)

Goal: Each subgroup should get a "positive outcome" at an equal rate.



Definition: $P(\hat{Y}|A=0) = P(\hat{Y}|A=1)$

where P denotes probability, $\hat{Y}$ is the predicted positive outcome, and A is the sensitive binary attribute that we use to split the groups.

Paid loan back

Did not pay loan back

9

Demographic Parity states that the proportion of each segment of a protected class (e.g. gender) should receive the positive outcome at equal rates.

So, if we have a similar example like above, but this time consider 2 groups, each with 4 members. If we set the threshold for the first group like that ---, then it means 4 out of 8 people got the loan. Hence, we need to set the threshold for the second group like that, such that they'll also have 50% chance to get a loan.

You see that their income threshold is lower. If we set it as high as for the other group, only 3 people would get a loan, and hence, both group would not get the positive outcome at the same rate.

Formal definition highlights that effect.

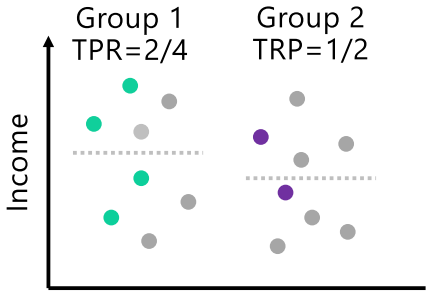
What can go wrong: for bank this is not ideal yet, since probably in the second group, the people do not have the right opportunity to pay that loan back. Imagine we had split that example by the group men and women, but we did not change some social structures. Image now the women want to travel the world, change career for something better and go back to university, but there is no social network in place to

pay them some form of salary while absent. Then, many of them will not be able to repay the loan and reinforce the historic stereotype that women should get less loans than men.

When should we use demographic parity?
•We want to **change the status quo** (e.g. more women get a loan)
•We are aware of **historical biases:** e.g. we might not have had a very representative group of women in the data in the past
•We have a **plan how to support the underrepresented groups.**
University admissions tend to look for demographic parity, giving more admission to students with difficult backgrounds. In this case, students have high academic potential, but their education might have been disrupted or they have had to overcome personal disadvantage. Students in this group can then usually receive extra support before beginning their degree, or get scholarships, mentoring or so.

Equal opportunity (group)

Goal: Each subgroup should get a "positive outcome" at an equal rate assuming the people in the group qualify for it.



Definition: $P(\hat{Y}|A = 0, Y = 1) = P(\hat{Y}|A = 1, Y = 1)$

where P denotes probability, $\hat{Y}$ is the predicted positive outcome, and A is the sensitive binary attribute that we use to split the groups, and Y is a true positive.

Paid loan back

Did not pay loan back

10

TPR=true positive rate (whether they should have been given the loan was accurately predicted: was predicted to pay it back and paid it back).
You get a loan with the same rate, no matter in what subgroup you are in if you qualify for a loan.

Based on the confusion matrix, we require the True Positive Rate (TPR) to be the same for each segment of the protected class.

What could go wrong? If we want to give out many loans in group 1, this means, they have a higher acceptance rate. Then, in turn, we also need to give out the loan to many people in group 2, and this can result in many false positives, i.e. many people will suffer from having a loan they cannot pay back. Hence, also the credit score would suffer more for people in group 2, leading to a disparate impact.

When should we use Equal opportunity:

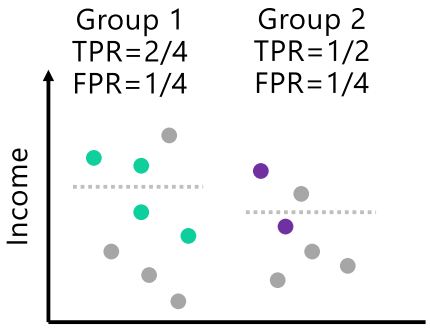
- When we really want to predict the positive outcome correctly very often (e.g. it is a good time, money is cheap and the bank wants to give out many loans).
- If a false positive is not costly. This is not the case for a loan (here both the

customer and the user suffer).

- A good application for equal opportunity is for example fraud detection (for credit cards). We want to notify every possible fraudulent transaction, and it is not so bad if we call the customer once more rather than once too few to verify a transaction.

Equalized odds (group)

Goal: Each subgroup should have the same TPR and FPR.



Definition: $P(\hat{Y}|A = 0, Y = y) = P(\hat{Y}|A = 1, Y = y)\ y \in \{0,1\}$

where P denotes probability, $\hat{Y}$ is the positive outcome, and A is the sensitive binary attribute that we use to split the groups, and y is the actual outcome.

Paid loan back

Did not pay loan back

11

Equalized odds is the most restrictive among the concepts discussed, so far.

Note that in all examples, whenever I said “exactly the same rate”, what I mean is that we optimize to get it to zero. We might not in every case be able to bring it to a complete zero gap. This can be well seen here. Imagine we just had one data point in group 2 more, then it would not fit. So we try to get as close to zero in the gap as we can.

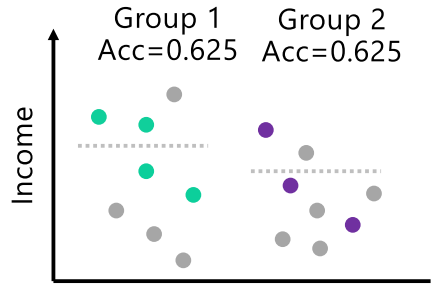
Given the restrictiveness of equalized odds, it might sometimes fail to produce highly accurate models.

When is it well suited?
- We care strongly about getting many positives, but also on minimizing the false negatives

This might be the best suited metric for a bank that wants to decide on who gets a loan. Since both the bank has interest in not having high false positives (they might loose money), and the users might suffer since their credit score suffers also for the future

Overall Accuracy & Treatment equality (group)

Goal: Overall accuracy: Each subgroup should have the same prediction accuracy.
Treatment accuracy: Look at errors that classifier makes rather than accuracy.



Paid loan back

Did not pay loan back

Definition Overall Accuracy:
$(Acc|\ A = 0) = (Acc|A = 1)$.

Definition Treatment Equality:
$(FN/FP|\ A = 0) = (FN/FP|A = 1)$.

Here Treatment equality:
Group 1: 2/1
Group 2: 2/1 → fulfilled

12

For “overall accuracy”, this is fulfilled if both groups have equal prediction accuracy – the probability of a subject from either positive or negative class to be assigned to its respective class. The definition assumes that true negatives are as desirable as true positives. In our example, this implies that the probability of an applicant with an actual good credit score to be correctly assigned a good predicted credit score and an applicant with an actual bad credit score to be correctly assigned a bad predicted credit score is the same for both male and female applicants.

FN=false negatives, FP=false positives.

Fairness through unawareness (individual)

A classifier satisfies this definition if no sensitive attributes are explicitly used in the decision-making process.

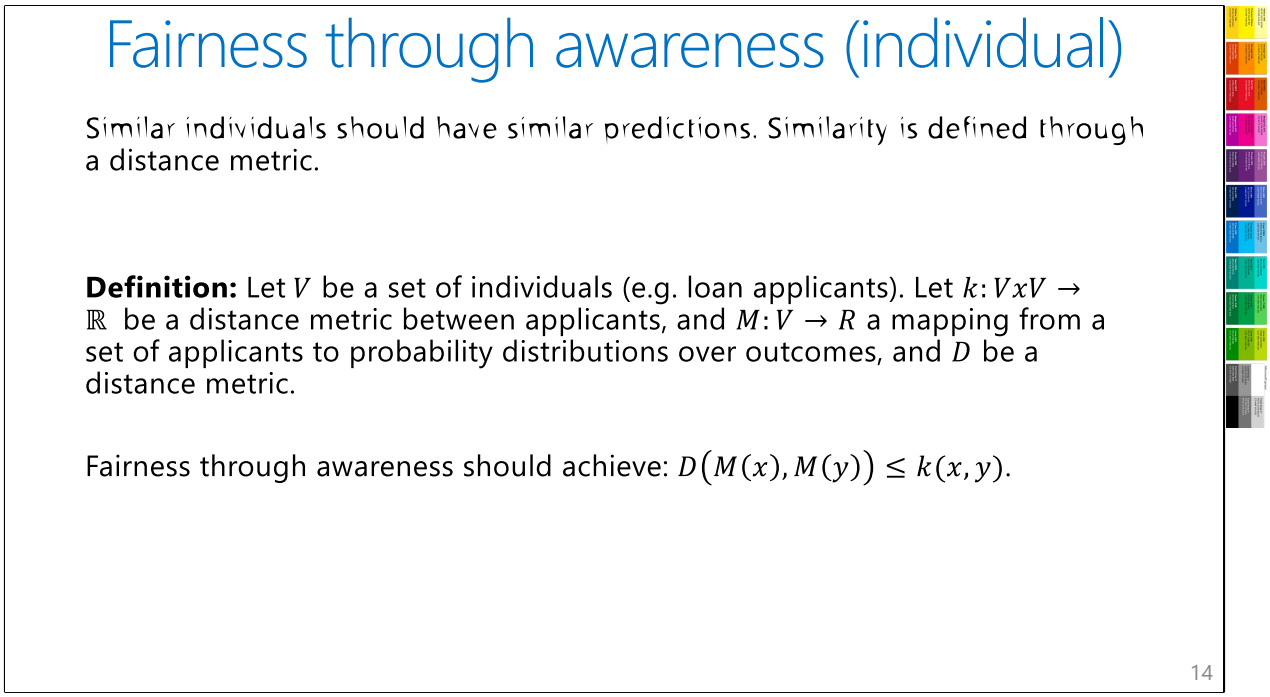
→ In our example: no gender-related feature can be used in training.

Definition: Let i and j be loan applicants whose features $X_i = X_j$ (apart from sensitive attributes) are identical, i.e. $X'_i = X'_j$.

Hence, $\hat{y}_i = \hat{y}_j$.

13

A classifier satisfies this definition if no sensitive attributes are explicitly used in the decision-making process. In our example, this implies that gender- related features are not used for training the classifier, so decisions cannot rely on these features. This also means that the classification outcome should be the same for applicants i and j who have the same attributes apart from the gender ones.



This definition is a more elaborated and generic version of the previous one.

Fairness is captured by the principle that similar individuals should have similar classification. The similarity of individuals is defined via a distance metric; for fairness to hold, the distance between the distributions of outputs for individuals should be at most the distance between the individuals.

For example, a possible distance metric k could define the distance between two applicants i and j to be 0 if the attributes in X (all attributes other than gender) are

identical and 1 if some attributes in X are different. D could be defined as 0 if the classifier resulted in the same prediction and 1 otherwise.

As another example, the distance metric between two individuals could be defined as the normalized difference of their ages: the age difference divided by the maximum difference in the dataset.

Intervention Types

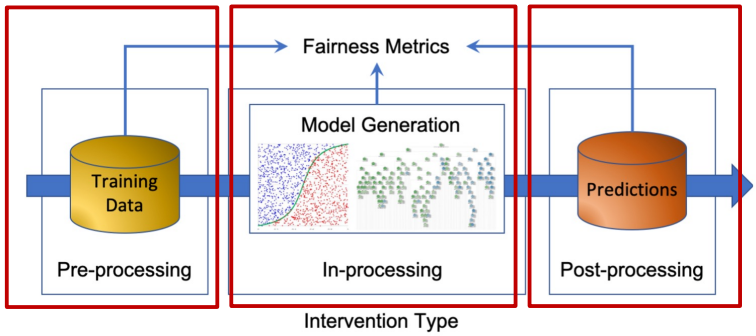


Figure taken from [1].

15

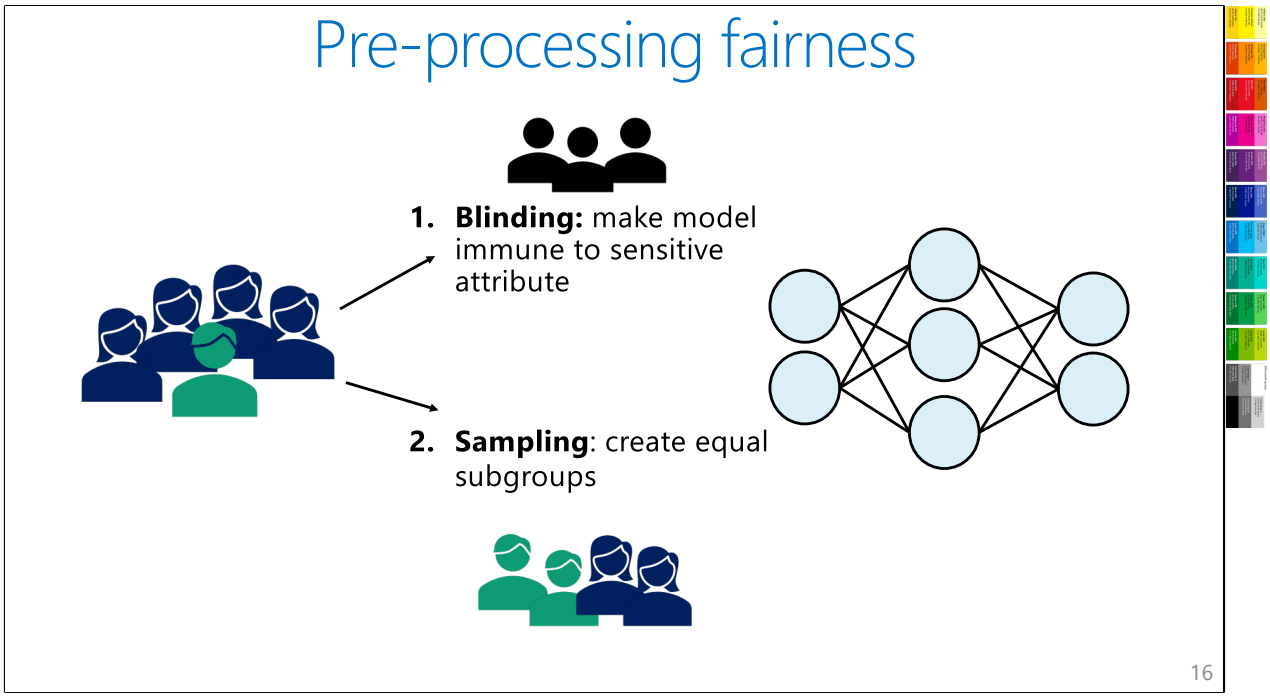
1.Pre-processing. Pre-processing techniques try to transform the data so that the underlying discrimination is removed. If the algorithm is allowed to modify the training data, then pre-processing can be used.
Pre-processing approaches recognize that often an issue is the data itself, and the distributions of specific sensitive or protected variables are biased, discriminatory, and/or imbalanced. Thus, pre- processing approaches tend to alter the sample distributions of protected variables or more generally perform specific transformations on the data with the aim to remove discrimination from the training data. The main idea here is to train a model on a "repaired" dataset. Pre-processing is argued as the most flexible part of the data science pipeline, as it makes no assumptions with respect to the choice of subsequently applied modeling technique.
2.In-processing. In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process. If it is allowed to change the learning procedure for a machine learning model, then in-processing can be used during the training of a model— either by incorporating changes into the objective function or imposing a constraint.
In-processing approaches recognize that modeling techniques often become

biased by dominant features, other distributional effects, or try to find a balance between multiple model objectives, for example, having a model that is both accurate and fair. In-processing approaches tackle this by often incorporating one or more fairness metrics into the model optimization functions in a bid to converge towards a model parameterization that maximizes performance and fairness.

3. **Post-processing.** Post-processing is performed after training by accessing a holdout set which was not involved during the training of the model. If the algorithm can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used in which the labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase.

Post-processing approaches recognize that the actual output of an ML model may be unfair to one or more protected variables and/or subgroup(s) within the protected variable. Thus, post-processing approaches tend to apply transformations to model output to improve prediction fairness. Post-processing is one of the most flexible approaches, as it only needs access to the predictions and sensitive attribute information without requiring access to the actual algorithms and ML models. This makes them applicable for black-box scenarios where the entire ML pipeline is not exposed.

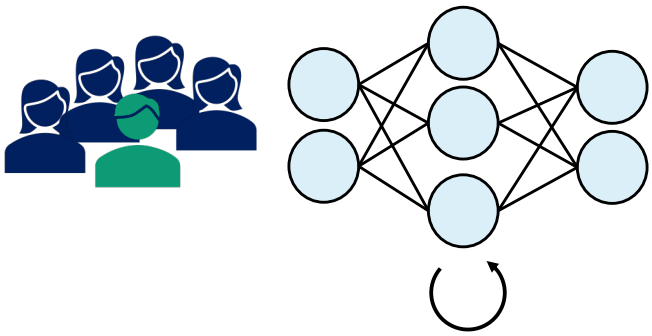
It can often be quite difficult to ascertain which type of approach will benefit a given scenario. A distinct advantage of pre- and post-processing approaches is that they do not modify the ML method explicitly. This means that (open source) ML libraries can be leveraged unchanged for model training. However, they have no direct control over the optimization function of the ML model itself. Yet, modification of the data and/or model output may have legal implications [17] and can mean models are less interpretable, which may be at odds with current data protection legislation with respect to explainability. Only in-processing approaches can optimize notions of fairness during model training. Yet, this requires the optimization function to be either accessible, replaceable, and/or modifiable, which may not always be the case.



- **Blinding:** Make a classifier "immune" to one or more sensitive variables. A classifier is, for example, race blind if there is no observable outcome differentiation based on race. Some work also calls the **omission** of sensitive variables from the training data as blinding. However, it might make sense to distinguish between immunity and omission of sensitive variables. While omission refers to not including the sensitive variables as input for the prediction models, immunity also considers the indirect effect that sensitive variables can have on other (input) variables of a prediction model. For instance, sensitive variables often are correlated with other variables in the data. In the US, for example, race is often correlated with the postal code, i.e. where people live. Hence, if we do not include the race attribute to the model for prediction (omission), the outcome might still be influenced by it indirectly if the postal code keeps on being an input. Approaches focusing on immunity aim to prevent these indirect effects from resulting in discrimination measured through the sensitive variable. Omission has been shown to decrease model accuracy and increase discrimination. In particular, it disregards the proxy variables. But also for immunity, it is usually hard to identify these proxy variables.

- **Sampling and Subgroup Analysis**: Sampling methods have two primary objectives: (1) to create samples for the training of robust algorithms, i.e., “correct” training data and eliminate biases; and (2) to identify groups (or subsamples) of the data that are significantly disadvantaged by a classifier where additional measures could be put into place… How are samples selected from a (large) dataset that is both diverse in features and fair to sensitive attributes? Unfortunately, without care, sampling can propagate biases within the training data, as ensuring diversity in the data used to train the model makes no guarantees of producing fair(er) models. A key challenge for sampling and subgroup analysis is to ensure that sufficient data are available for each subgroup. Otherwise, this method of sampling can negatively affect performance. Outliers can also be problematic.

In-processing fairness



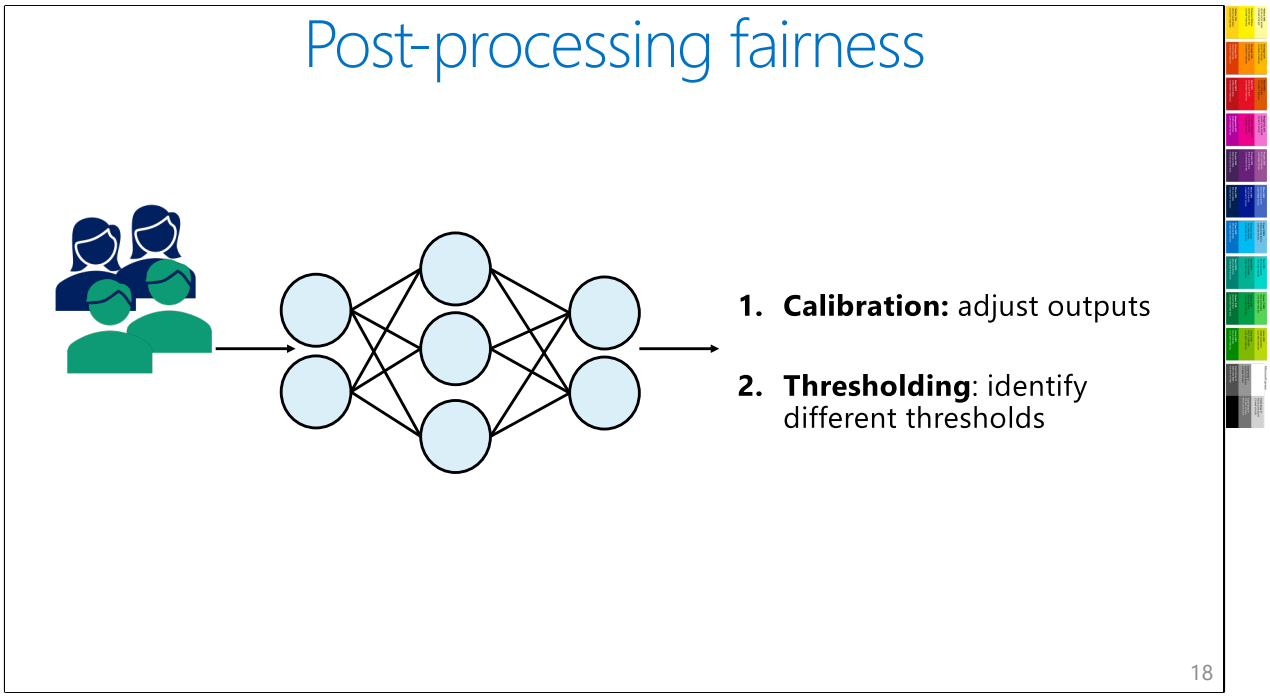
1. **Reweighting:** change the impact of (individual) instances

2. **Regularization / Constraint Optimization**: include fairness into training

17

- **Reweighting** (can also count still as pre-processing): Change the "impact" of instances (observations) on the prediction model during training to promote "fair(er)" handling of sensitive variables and/or underprivileged groups. Weights can be introduced for multiple purposes: (1) to indicate a frequency count for an instance type, e.g. if you have 100 samples for one class, but only 50 for another, you might just weight the impact of these 50 double to come to the same number. (2) to place lower/higher importance on "sensitive" training samples. I.e., if you already know you have to little samples with a particular set of attributes, you can just weight up the few that you already have. Reweighing subtly changes the data composition, making the process less transparent which is a disadvantage if you want to explain how you came to a given model – we'll cover explainability in the next lecture.
- **Regularization and Constraint Optimization**: Extend the classifier's loss function such that it penalizes "unfair" outcomes. Classically, regularization penalizes the complexity of the ML model to inhibit over-fitting. Applied to fairness, regularization means adding penalty terms to penalize the classifier for discriminatory practices. When extending the classifier's (convex) loss function with fairness terms, researchers

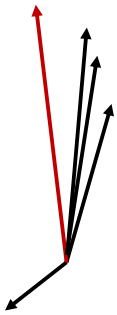
typically seek to balance fairness and accuracy. Constraint optimization approaches often include notions of fairness7 in the classifier loss function operating on the confusion matrix during model train- ing. often approaches for fair ML are not stable, i.e., subtle changes in the training data significantly affect performance (comparatively high standard deviation).



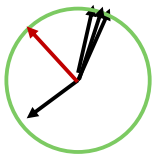
- **Calibration**: To adjust the probability outputs of a model such that the portion of predicted positive outcomes matches that of positive examples across (or within) all (sub)groups in the dataset. Calibration is the process of ensuring that the proportion of positive predictions is equal to the proportion of positive examples . In the context of fairness, this should also hold for all subgroups (protected or otherwise) in the data. Calibration is particularly useful when the output is not a direct decision but used to inform human judgment when assessing risks (e.g., awarding a loan), So what you can do is to observe over the course of your model predictions to how many women a loan was given and to how many men. You might then manually give more loans to women where the model actually was negative. Calibrating an ML model for multiple protected groups and/or using multiple fairness criteria at once has been shown to be difficult (impossible).
- **Thresholding**: Thresholding methods try to find classification thresholds for each group, such that fairness constraints are met . This is like in the examples for the fairness metrics we saw. If the only decision variable is the income, for some groups, we might already warrant a loan at a lower income than for others. Or if the model outputs a confidence score for a

positive prediction, if you say ‘yes’ with confidence 0.8 for one group, you could also already say yes for confidence ‘0.7’ for another group. The underlying idea here is to incentivize good performance (in terms of both fairness and accuracy) across all classes and groups. Thresholding can claim compelling notions of equity, however, only when the threshold is correctly chosen. The main challenge for thresholding approaches is to find a tolerance level for unfairness in the calculation of threshold values.

Differential privacy reduces accuracy



Original Gradient



Clipped Gradient

19

Why is this the case: High level: the less signal you have and then add the same amount of noise, the lower your signal to noise ratio, the less signal you can learn from. One particular aspect is also the Clipping. Clipping can alter the gradient direction. The gradient is the aggregate over all existing gradients. Hence, if some point to the right direction, and then a few to the wrong. As long as the right ones are larger or more, there is no problem, but when they are smaller, or when there are fewer, then the influence from the “outlier” becomes stronger.

This is particularly severe towards the end of training. Then, you can imagine that the model is already very good for all “inliers”, i.e. standard data points. This means that their gradients are near zero. Only for the outliers, the model is not good yet. Hence, only for them, it would have large gradients, and then it can easily go into the wrong direction.

Differential privacy reduces fairness

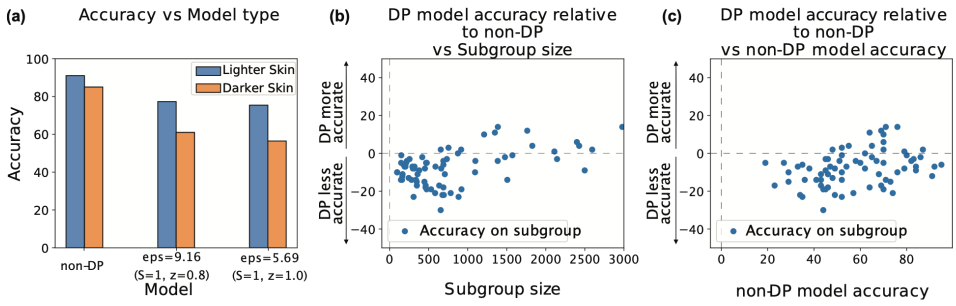


Figure 1: Gender and age classification on facial images.

Figure taken from [5].

20

DP in its standard application with DPSGD or PATE is unable to learn too much information about the tails of a data distribution, resulting in a loss of accuracy that can disproportionately affect small groups. (hence, DP reinforces biases and makes them even stronger).

In addition to the clipping which affects the gradient direction, also noise addition can be bad.

We see results from a paper that measures accuracy gap when training with DP. We see that the higher the epsilon (i.e. the better the privacy), the greater the accuracy difference between the prediction for lighther vs. darker skin. (a)

We see also that in particular, the smaller a subgroup, the more the accuracy drop. (b) This means, smaller groups are much more affected by private training.

Accuracy drop is strong over all final model utilities (c).

Why would we see a similar picture in PATE? Well, if all teachers have one data

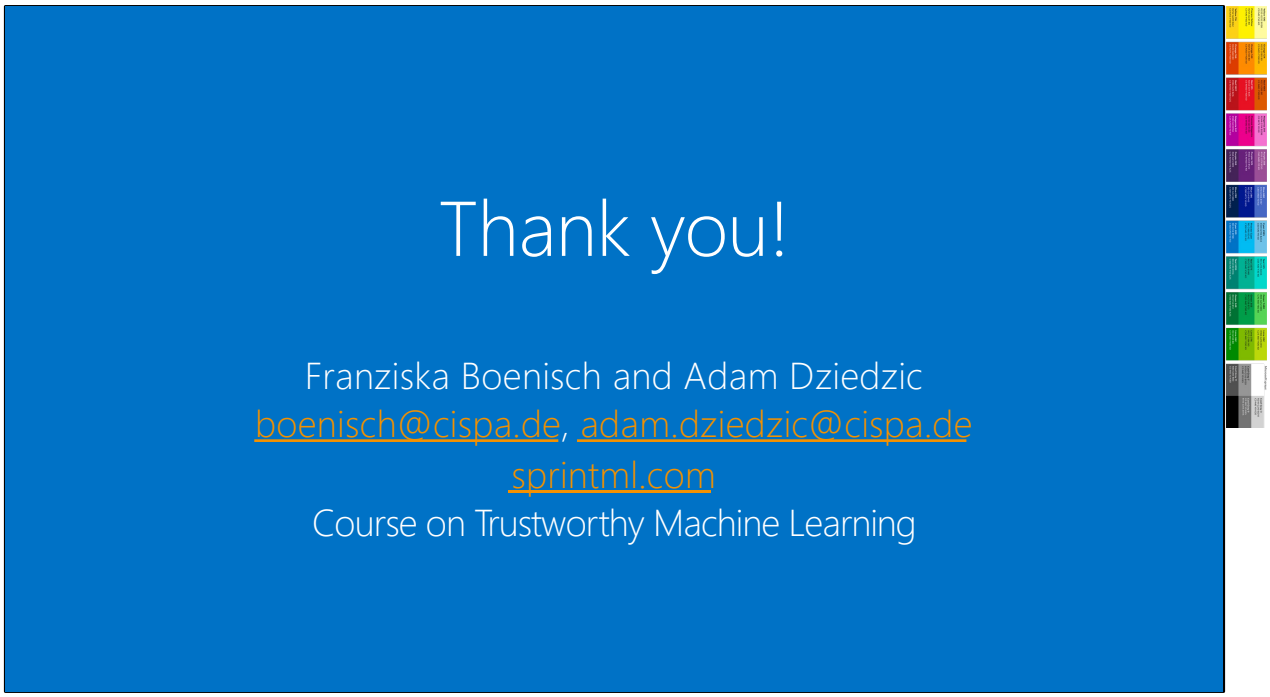
point (or even none) from a subgroup, then they cannot learn about it, and hence, not predict correctly. If we were to give all data from a subgroup to the same teacher, such that at least they could learn about that group, we would face the issue that when we aggregate the teachers' votes, the one teacher has only one voice, and hence, would not have a great influence on the final label.

Challenges and open problems

1. **One notion of fairness**: many existing (and partially contradicting notions of fairness.
2. **Identifying unfairness**: given a dataset or an ML model, find directly whether there might be problems with fairness.
3. **Trade-offs:** achieve fairness while preserving other desirable aspects, such as accuracy, privacy, explainability.
4. **From equality to equity:** give every group the resources they need to succeed → goes beyond ML fairness.

21

1.Synthesizing a definition of fairness. There are many (partially opposing) definitions of fairness. The „correct one“ should be chosen according to the application. However, it is nearly impossible to understand how one fairness solution would fare under a different definition of fairness
2.Searching for Unfairness. Given a definition of fairness, it should be possible to identify instances of this unfairness in a particular dataset.
3.Trade-offs (e.g. fairness decreases accuracy, and if we want privacy, we degrade fairness
4.From Equality to Equity. The definitions presented in the literature mostly focus on equality, i.e. ensuring that each individual or group is given the same amount of resources, attention or outcome. However, little attention has been paid to equity, which is the concept that each individual or group is given the resources they need to succeed —> this is a challenge in society as well



With that, thank you very much for your attention.

Further Reading

[1] Caton, Simon, and Christian Haas. "Fairness in machine learning: A survey." *ACM Computing Surveys* 56, no. 7 (2024): 1-38.
[2] Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning." *ACM computing surveys (CSUR)* 54, no. 6 (2021): 1-35.
[3] Verma, Sahil, and Julia Rubin. "Fairness definitions explained." In *Proceedings of the international workshop on software fairness*, pp. 1-7. 2018.
[4] Suriyakumar, Vinith M., Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. "Chasing your long tails: Differentially private prediction in health care settings." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 723-734. 2021.
[5] Bagdasaryan, Eugene, Omid Poursaeed, and Vitaly Shmatikov. "Differential privacy has disparate impact on model accuracy." *Advances in neural information processing systems*32 (2019).
[6] Esipova, Maria S., Atiyeh Ashari Ghomi, Yaqiao Luo, and Jesse C. Cresswell. "Disparate Impact in Differential Privacy from Gradient Misalignment." In *The Eleventh International Conference on Learning Representations*. 2022.

23