Seminar Thesis

FairML and the SQF dataset

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, February, 24th 2025

Submitted in fulfillment of the requirements for the degree of B. Sc.
Supervised by FairML and the SQF dataset

Abstract

In the first half of this paper we provide an introduction to the most common metrics and methods in fair machine learning. We then apply the theoretical concepts to the New York Stop, Question and Frisk dataset, which will showcase difficulties that come with fairness in practice. This leads us to explore the problem of selection bias and related issues. We turn our focus to studies that have worked with the SQF dataset and established interesting theoretical results; residual unfairness, bias reversal and bias inheritance. Value of this paper: compare and contrast traditional fairness metrics, use them in a real world setting, show its limitations in this setting, present how they are adressed in more advanced ways and try to explain why traditional metrics can not reflect the whole situation.

Contents

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A \vert Y$	$Y \perp A \vert \hat{Y}$

1 Introduction

2 Related Work

3 Fairness Metrics

When one starts to get into the topic of fairness in machine learning, it is easy to get overwhelmed by the sheer amount of definitions and metrics that are out there. In this chapter we try to group them in an intuitive way and motivate them in the hope to bring some clarity to readers. It is helpful to group fairness metrics in the following ways.

1. Group fairness vs. individual fairness
2. observational vs. causality-based criteria

Broadly speaking, group fairness aims to create equality between groups and individual fairness aims to create equality between two individuals within a group. Observational fairness metrics act descriptive and use the observed distribution of random variables characterizing the population of interest to assess fairness while causality-based criteria make assumptions about the causal structure of the data and base their notion of fairness on this. On the basis of these fundamental ideas, a plethora of formalizations have emerged. Most of them concern themselves with defining fairness for a binary classification task and one binary protected attribute (PA). The extension to a multiclass PA is the easiest. The extension to multiple sensitive attributes, on the other hand, brings challenges with it. Also, the extension from binary classification to other tasks, such as neural networks, LLMs and other models is subject of ongoing research. As this work is meant to help you start thinking about fairness in machine learning, we will limit ourselves to the binary classification case. To bring more clarity we will illustrate the general fairness metrics already in the context of the SQF dataset.

3.1 Group fairness

The notion of fairness underlying group metrics is that discrimination of certain groups of the population defined via the PA should be prevented. The groups metrics presented here are observational metrics. They can be separated into three main categories, independence, separation, and sufficiency.

Independence

Independence is in a sense the simplest group fairness metric. It requires that the prediction $\hat{Y}$ is independent of the protected attribute A, so $\hat{Y} \perp A$. This is fulfilled when for each group the same proportion is classified as positive by the algorithm. In other words,

the positive prediction ratio (ppr) should be the same for all values of A. For a binary classification task with binary sensitive attribute this can be formalized as
demographic parity/statistical parity

$$P(\hat{Y}|A = a) = P(\hat{Y}|A = b)$$

This indeed seems like a simple, perhaps too simple definition of fairness, as it only looks the distribution of $\hat{Y}$ conditioned on the group membership. There are extensions of this idea, such as conditional statistical parity. This metric allows to condition on A and a set of legitimate features E. For instance, predictive parity would mean that we require equal prediction ratios between PoC and white people while conditional statistical parity requires equal prediction ratios between PoC and white people who *live within the same borough of New York* (E = borough). This can be seen as a more nuanced approach, as it allows tacking additional information into account. The other two groups of group fairness metrics, Separation and Sufficiency can both be derived from the error matrix.

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	TN	FN
$\hat{Y} = 1$	FP	TP

Separation

Separation requires independence between $\hat{Y}$ and A conditioned on the true label Y, so $\hat{Y} \perp A|Y$. This means that the focus is on equal error rates between groups, which gives rise to the following list of fairness metrics:

- Equal opportunity/ False negative error rate balance

  $$P(\hat{Y} = 0|Y = 1, A = a) = P(\hat{Y} = 0|Y = 1, A = b)$$

  or $P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b)$

- Predictive equality/ False positive error rate balance

  $$P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = b)$$

  or
  $P(\hat{Y} = 0|Y = 0, A = a) = P(\hat{Y} = 0|Y = 0, A = b)$

- Equalized odds

  $$P(\hat{Y} = 1|Y = y, A = a) = P(\hat{Y} = 1|Y = y, A = b) \forall y \in \{0, 1\}$$

- Overall accuracy equality:

  $$P(\hat{Y} = Y|A = a) = P(\hat{Y} = Y|A = b)$$

- Treatment equality:
$$\frac{\text{FN}}{\text{FP}}\Big|_{A=a} = \frac{\text{FN}}{\text{FP}}\Big|_{A=b}$$

Equal opportunity requires the false negative rates, the ratio of actual positive people that were wrongly predicted as negative, to be equal between groups. Therefore, it is also called false negative error rate balance. When there false negative rates are equal between groups, then the true positive rates between groups are also equal. This means requiring equal false negative rates or equal true positive rates between groups results in the same effect. Predictive equality follows the same principle as equal opportunity but instead of focusing on the false negatives, it focuses on the false positives. Again, if a classifier has equal false positive rates between groups, it also has equal true negative rates. Equalized odds combines equal opportunity and predictive equality. It requires that the false positive and true positive rates are equal between groups, and is in this sense stricter than either of them alone.
In itself, these error rates are detached from the context of fairness and used in general in machine learning to assess the performance of a classifier. In essence the group metrics we outlined so far do nothing other than picking a performance metrics from the confusion matrix and requiring it to be equal between two (or more) groups in the population. This means the well-known trade-offs for example between false positive and true positive rate are also present in the fairness metrics. As more people get correctly classified as positive usually also more people get wrongly classified as positive. Source With this comes the difficulty to choose "the right" metric for the specific task. In general one can think about this in the same way as when choosing a performance metric for a binary classifier. In setting in which a positive prediction leads to a harmful outcome, as in the SQF setting, it often makes sense to focus on minimizing the false positive rate, while a higher false negative rate is accepted as a trade-off. This argumentation follows the idea of. The authors distinguish between punitive and assistive tasks to help choose the right fairness metric. For punitive tasks metrics that focus on false positives, such as predictive equality are more relevant. For assistive tasks, such as deciding who receives some kind of welfare, a focus on minimizing the false negative rate could be more relevant, so equal opportunity would be more suitable.

Sufficiency

Sufficiency requires independence between Y and A conditioned on $\hat{Y}$, so $Y \perp A|\hat{Y}$. Intuitively this means that we want a prediction to be equally credible between groups. When a white person gets a positive prediction the probability that it is correct should be they same as for a black person. This leads to the following fairness metrics:

- Predictive parity/ outcome test requires that the probability of actually being positive, given a positive prediction is the same between groups.
$$P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b)$$
- Equal true negative rate follows the same principle as predictive parity. It requires that the probability of actually being negative, given a negative prediction is the

same between groups.:

$$P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A = b)$$

- If we instead look at errors again, we can require equal false omission rates:

$$P(Y = 1|\hat{Y} = 0, A = a) = P(Y = 1|\hat{Y} = 0, A = b)$$

- Or equal false discovery rate:

$$P(Y = 0|\hat{Y} = 1, A = a) = P(Y = 0|\hat{Y} = 1, A = b)$$

- Conditional use accuracy equality:

$$P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b) \wedge P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A =$$

False omission describes the case in which an actual positive person is predicted as negative and can be highly relevant in assistive settings, such as description of a medical treatment. False discovery rate describes the case in which an actual negative person is predicted as positive. This should be taken into account in punitive settings, in which we do not want to convict innocent people. Just as for equalized odds as Separation metric, we can build a stronger Sufficiency metric requiring multiple conditions to hold simultaneously, e.g. conditional use accuracy equality. Hopefully, the pattern becomes clear now. While it is easy to get overwhelmed by the amount of definitions at first, taking a closer look, it becomes clear that they are constructed in a structured way. In fact, equal false omission rate and equal false discovery rate were not introduced in the paper Verma and Rubin 2018 but are implemented in `mlr3fairness`, and it is clear that they follow the same pattern as the other metrics.

Most (binary) classifiers work with predictions scores and a hard label classifier is applied only afterwards in form of a threshold criterion. It should therefore come as no surprise that instead of formulating fairness with $\hat{Y}$ there exist fairness metrics that use the score S, which typically represents the probability of belonging to the positive class. Instead of conditioning on $\hat{Y}$ as Separation metrics, we can simply condition on S and define:

Calibration:

$$P(Y = 1|S = s, A = a) = P(Y = 1|S = s, A = b)$$

Calibration requires that the probability for actually being positive, given a score s is the same between groups. So the idea is a more fine-grained version of predictive parity. As the score can usually take values from the whole real number line, this can in practice be implemented by binning the scores Verma and Rubin 2018.

To compare the group fairness criteria, sufficiency takes the perspective of the decision-making instance, as usually only the prediction is known to them in the moment of decision. For example, the police, who do not yet know the true label at the time when they are supposed to decide whether someone would become a criminal. As separation criteria condition on the true label Y it is suitable when we can be sure that Y is free from any bias, so to say when Y was generated via an objectively true process (this will become clearer in the chapter on bias). Independence is best, when we want to enforce a form of equality between groups, regardless of context or any potential personal merit. While this seems to be useful in cases in which the data contains complex bias, it is unclear whether these enforcements have the intended benefits, especially over the long term. Reference?. It is good to understand the difference in perspectives each of the group fairness metrics take, because many of them cannot be satisfied simultaneously. This is known as the impossibility theorem Hardt et al. 2016. This means one has to decide on either Independence, Separation or Sufficiency and the choice should fit the context of the data and the decision-making process. Lastly, we note that these are not all the group fairness metrics that exist, but broadly speaking other metrics are variations of the presented ones. Some more metrics are listed in the appendix.

3.2 Individual fairness

If we want to equalize e.g. the false positive rates between two groups and currently group a has a higher false positive rate than group b, this would lead us to lowering the prediction threshold for b, such that more actual negative people would get classified as positive. Or if we would need to set a higher threshold for group a, such that it becomes harder for them to be classified as positive. Depending on the context, either option can seem unfair. So by trying to equalize a given metric between groups, it can happen that individuals within a group are treated unequally. Individual metrics therefore shift the focus. The underlying idea of fairness is that similar individuals should be treated similarly.

Fairness through awareness (FTA)

FTA formalizes this idea as Lipschitz criterion.

$$d_Y(\hat{y_i}, \hat{y_j}) \leq \lambda d_X(x_i, x_j)$$

d_Y is a distance metric in the prediction space, d_X is a distance metric in the feature space and λ is a constant. The criterion puts an upper bound to the distance between predictions of two individuals, which depends on the features of them. In other words, if two people are close in the feature space, they also should be close in the prediction space. The challenge of FTA is the definition of the equality in the feature space. Castelnovo et al. 2022 states ”defining when two individuals are similar is not much different from defining fairness in the first place”. In the SQF context, it could make sense to define similar individuals based on yearly income, age and neighbourhood. Yet one could easily argue that taking the criminal history into account is important as well. After the decision for a legitimate set of features has been made, the next challenge is to choose a distance metric that appropriately captures the conceptual definition of similarity defined via the selected features. FTA does not have one clear solution and requires domain knowledge and the choice of d_X should take context-specific information into account.

Fairness through unawareness (FTU) or blinding

In contrast to FTA, blinding should give a simple, context-independent rule. It tells us to not use the protected attribute explicitly in the decision-making process. When training a classifier this means discarding the PA during training. Since FTU is a more procedural rule than a mathematical definition, there exist multiple ways to test whether the blinding worked for a classifier. One approach is to simulate a doppelgänger for each observation in the dataset. This doppelgänger has the exact same features except the protected attribute, which is flipped. If both these instances have the same prediction, the algorithm would satisfy FTU Verma and Rubin 2018. [1] Other ways to assess FTU can be found in Verma and Rubin 2018. A problem blinding has been proxies. These are variables that are strongly correlated with the sensitive attribute. It is not enough to simply mask the information of the sensitive attribute during training because discrimination can persist via these proxies. For SQF this would mean that we remove the race attribute during training. A person's ethnicity, however, is strongly correlated with their place of residence. Thus, indirect discrimination based on ethnicity remains, even though the information was not directly available during training. **Suppression** extends the idea of blinding and the goal is to develop a model that is blind to not only the sensitive attribute but also the proxies. The drawback is, that it is unclear when a feature is sufficiently high correlated with the sensitive attribute to be counted as proxy. Additionally, we could lose important information by removing too many features Castelnovo et al. 2022.

3.3 Causality-based fairness metrics

In contrast to observational fairness metrics, causality-based notions ask whether the sensitive attribute was the *reason* for the decision. If a certain (harmful) decision was made *because of* the value of the sensitive attribute of a person, we deem the algorithm as unfair.
Group-level: FACE, FACT (on average or on conditional average level) (Zafar et al. 2017)
Individual-level: counterfactual fairness, path-based fairness (Kusner et al. 2017) The two most common individual fairness metrics are counterfactual fairness and path-based fairness.

3.4 Comparison and Summary

The difference between observational and causal clear, really different approach. The division in group and individual fairness metric actually more of a nuanced differentiation. The observational metrics can rather be ordered on a plane, depending on how much information of the situation via other features X they allow. Traditional group metrics like demographic parity, equal error rate metrics and sufficiency metrics only work with the distribution of $Y, \hat{Y}, X, A$. The individual fairness metrics take more information of the non-sensitive feature into account in order to define similarity. Metrics such as

[1] This can be seen as a from of FTA, in which we chose the distance metric to measure a distance of zero only if two people are the same on all their features except for the protected attribute. In this special case FTA and FTU are measured in the same way.

conditional demographic parity lie in between, as we allow for a relevant subset of non. Sensitive feature to be part of the definition. Castelnovo et al. 2022 therefore depict this as a plane. The amount of approaches to measure fairness shows the complexity of the topic. There is not *the* right fairness metric to choose, but there can be the best one depending on the context and the data. The next section will present ways to digitate algorithmic bias once detected by one of the fairness metrics.

4 Fairness Methods

Fairness methods

Another question fair machine learning deals with is how algorithms can be adjusted such that they fullfil one of the above fairness metrics. Depending on when they take place in the machine learning pipeline, we distinguish between preprocessing, inprocessing or postprocessing methods. Preprocessing methods have the idea that the data should be modified before training, so that the algorithm learns on ”corrected” data. Reweighing observations before training is an example for a preprocessing method, that we will use in our case study in chapter x. In Processing methods modify the optimisation criterion, such that a it also accounts for a chosen fairness metric. Introducing a regularization term to the loss function is one example of such modifications. Postprocessing methods work with black box algorithms, just like preprocessing methods. We only need the predictions from the model to adjust them so that again a chosen fairness metric is fullfilled. One example for this is thresholding, where we set group specific thresholds to re-classify the data after training. We will discuss a post-processing approach by Hardt et al. 2016 in a following chapter.

Bias

We want to end this general introduction into fair machine learning by outlining the context in which the algorithm is usually embedded. On this note we also advice practitioners to think about the source of bias that could be present in your situation, as this *should* influence how fairness is defined and what fairness adjustments are appropriate. This will motivate the potential difficulties that can arise when implementing fairness in the real world. Caton and Haas 2024 describe the situation as follows. The algorithm is embedded in a feedback loop with the user and data. We as a society make decision, which reflect our reality. We make our reality measurable by collecting data. The algorithm learns from this data and makes predictions, on which we base new decisions. At each of these three points bias can be introduced into the process and, above all, bias can also be reinforced in the course of this process. In the context of the Stop, Question, and Frisk data, historical bias and selection bias are probably the most relevant sources of bias. Historical bias can shows itself in different ways. In our case it would mean that we assume that some people in our data have repeatedly experienced discrimination in terms of being arrested. Selection bias refers to the fact that the data is not representative of the population of New York City, because the decision to stop someone is based on a biased decision policy.

5 Case Study: Stop, Question, and Frisk

Stop, Question, and Frisk data

The legal sector is one in which Arms have been deployed, more often than not accompanied by public debate and protests about targeted policing and racial discrimination (COMPASS as the most popular example). We will turn our focus to the stop, question, and frisk (SQF) dataset published by the New York police department (NYPD). A. Fabris and S. Messina and G. Silvello and G.A. Susto scanned more than x datasets to diversify the datasets that are used in the fairness literature. They recommend it as suitable dataset for fairness research. First, we will give some context to the dataset. We will continue with descriptive analysis and finally examine fairness of an algorithm trained on this data.
Since x the stop, question, and frisk practice is implemented in New York City. A police officer is allowed to stop a person if they have reasonable suspicion that the person has committed, is committing, or is about to commit a crime. During the stop the officer is allowed to frisk a person (pat-down the person's outer clothing) or search them more carefully. The stop can result in a summon, an arrest or no further consequences. After a stop was made, the officer is required to fill out a form, documenting the stop. This data is published yearly by the NYPD. Many citizens have criticized the stop and frisk practice. There is disagreement about whether the strategy is effective in reducing the crime rates of the city cite some studies. The police has been repeatedly criticized for over-targetting people of colour. Stop and Frisk practice during 2004 to 2012 has been deemed as unconstitutional. Source

Data description

For our analysis we look at the stops from 2023 as they were the most recent recordings at the time of writing this paper. The raw 2023 dataset consists of 16971 observations and 82 variables. We first discarded all the variables that have more than 20% missing values, which leaves 34 variables. From this reduced dataset we filter out the complete cases and end up with 12039 observations. [2] From the clean dataset we choose x variables as features that contain information about the stop and demographic information of the stopped person. Table x gives an overview of them. We choose the arrestment of a suspect as target and the race as protected attribute. For the fairness audit later in the chapter we dichotomize the PA to adjust our situation to the common binary classification, binary PA scenario in the fairness literature. For a more nuanced descriptive analysis we only summarize "Black Hispanic" and "Black" into the group "Black" and "American Indian/ Native American" and "Middle Eastern/ Southwest Asian" into the "Other" category. In the cleaned 2023 data about 31% of stops result in an arrest. Overall racial disparities in arrestment rates are low. The arrestment rate for white suspects is the highest. Figure 3. As group fairness metrics are observational and constructed from the joint probability

[2] Simply discarding the missing values and only training on complete cases is discouraged by Fernando et al. 2021. We opt for this approach regardless, since imputation of the missing values is not straight forward but treating missing values as an extra category (which some random forest learners in mlr3 can do) will introduce complications when we implement some fairness methods later on.

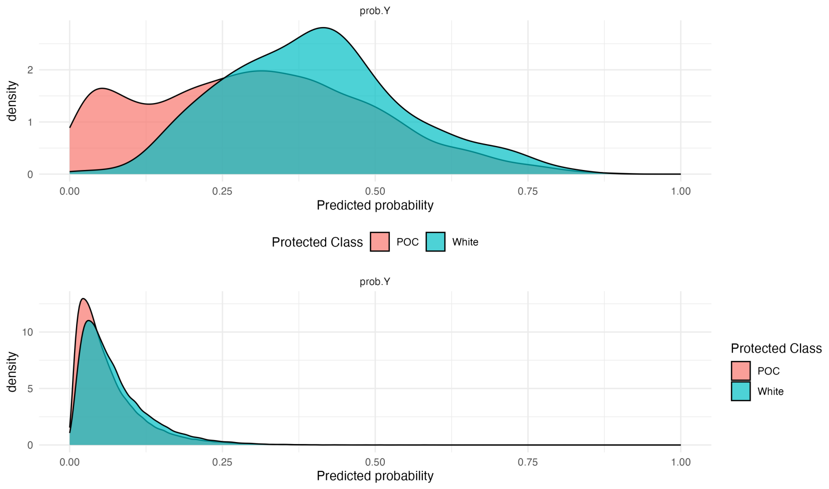

Figure 1: Density of predicted probabilities both groups.

of $Y, \hat{Y}, A$ alone, this already gives us a hint that the classifier might show little racial disparities.

Fairness Auditing

To train a random forest classifier on the 2023 data to predict the arrest of a person, we dichotomize the race attribute by grouping "Black" and "Hispanic" as people of colour ("PoC") and "White", "Asian", and "Other" as white ("White"). We select specific features that should resemble the information that were available to the officer at the time of the stop. Additionally, we control for variables, such as the time of the stop or whether the officer was wearing a uniform. This selection of features can be similarly found in previous studies of SQF. With many of the group fairness metrics implemented in mlr3fairness, we can measure the (group) fairness of our models. As suspected after data description, we find that the random forest classifier is already fair. There are minor differences between groups, but exact equality cannot be expected. It is common to allow for a certain margin of error ϵ in practice. Especially the error rates (FNR, FPR) are very similar between groups, thus Separation seems to be satisfied overall. Sufficiency metrics have larger differences, though they are still minor. Mlr3fairness offers functions to easily visualize fairness. First, in Figure 1 we plot the density of risk scores output by the random forest for each group. The risk score represents the probability of getting a positive prediction $P(\hat{Y} = 1)$, which is undesirable in the SQF context. White subjects tend to have higher predicted probabilities of being arrested, their mode lies around 0.15 while the mode of the PoC group is around 0.05. This means we have more low-risk PoC in the data. Next, from Figure 2 we can see that the positive predictive value (Sufficiency) has a relatively large difference between groups, while the false positive rate is practically the same between groups (Separation). It comes to no surprise that equalized odds, which is based on error rates, is satisfied. Finally, the accuracy between groups is not as equal as the error rates, but the absolute difference is still smaller than 0.05, which is a common ϵ to choose. Given that the classifier is fair from a group perspective, it does not make

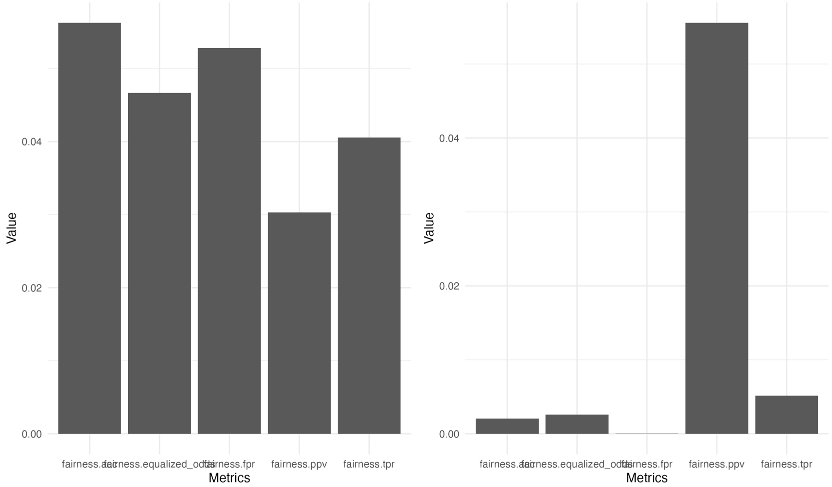

Figure 2: Comparison of fairness metrics.

sense to experiment with any of the implemented fairness methods in mlr3. At most, we could try to address the disparities in sufficiency metrics, but the common methods in the package are designed to address concerns with independence or separation.

Fairness Experiment

Inspired by the chapter on fairness in the mlr3book, we decide to experiment with some fairness methods for the SQF data. `Mlr3fairness` currently has two preprocessing methods, one post-processing method and some fairness adjusted models implemented. We decide to use a reweighing methods that works with assigning weights to the observations to equalize the distribution of $P(Y|PA)$. The in processing method is a fairness-adjusted logistic regression implemented in `mlr3fairness` inspired by Zafar et al. This method optimizes for statistical parity (independence). The post-processing method we choose aims for equalized odds, and it works by randomly flipping a subset of predictions with pre-computed probabilities in order to satisfy equalized odds constraints.
It is more interesting to compare the classifier trained on data that comes from the unconstitutional period 2004 to 2012. We decide for 2011 as it is the year with the most stops. We carry out the same data cleaning steps for the 2011 data as before, starting with 685724 recorded stops and reducing this to 651567 clean observations. Note, these are more than 50 times more stops than in 2023. The 2011 data has substantially more low-risk stops, only around 6% of stops result in an arrest. This is a stark contrast to the 2023 data, where 31% of stops result in an arrest. The differences in arrestment rate between groups are slightly lower for 2011 and the highest arrestment rate remains to be for the white group. Figure 3 Due to the large proportion of low-risk stops in 2011 the predicted probabilities are generally low. The x-axis is cut-off at a probability of 0.1, otherwise it would be hard to see anything as the vast majority of probability mass lies in the small regions. The measured racial disparities are interestingly not greater than in 2023. We see that for the classifier trained on 2011 stops, equalized odds has the greatest

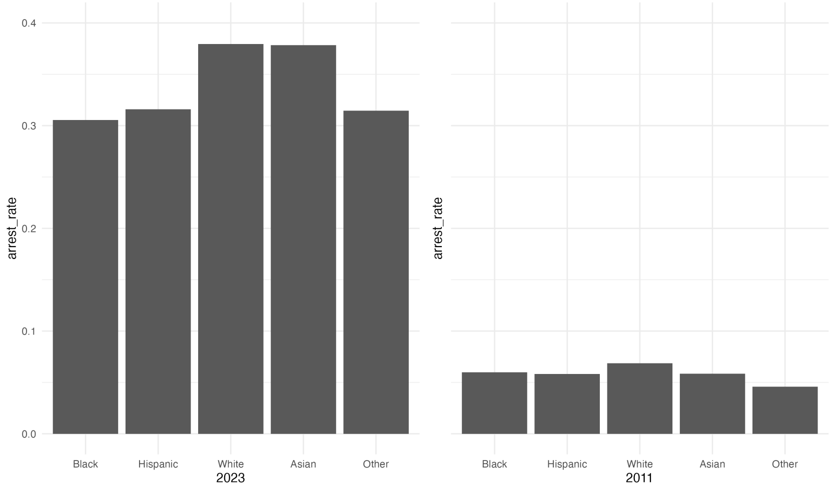

Figure 3: Comparison of arrestment rates for 2023 (left) and 2011 (right).

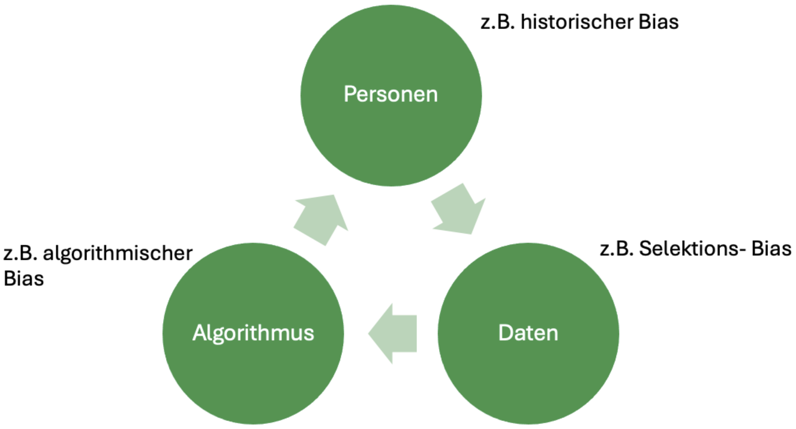

Figure 4: The bias loop.

difference, but overall the differences in predictions rates across groups are very small. Our fairness audit did not show any substantial disparities in fairness metrics. Does this mean the classifier is fair? It is easy to come to such conclusions, especially if fairness is not the major concern of the practitioners but more of a nuisance criterion that should be fulfilled. However, to truly ensure a fair practice, it is crucial to look at the context in which the algorithm is embedded.

Bias and the feedback loop

Usually fairness is a concern in the first place, because the algorithm should be implemented as an ADM to assist decision-making in some way. As such it could influence if someone gets admitted to college, gets a loan or is released from prison. The algorithm does not exist in isolation, but is embedded in a loop with data and the user. We make the circumstances of a decision measurable by collecting data. The algorithm learns from

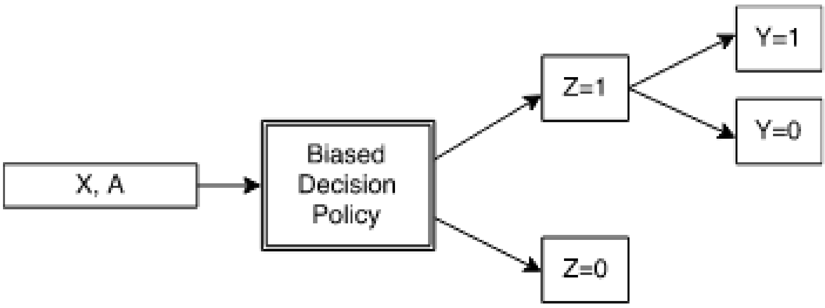

Figure 5: Selection bias in the SQF data.

this data to make an optimal prediction, on which the decision-makers base their judgement on Figure 4. At each step of this loop, bias can be introduced in the process and, more dangerous, be amplified as the algorithm influences decision-making on a large scale. This means that every fairness project comes with the task to understand where the data comes from and how exactly the algorithm will be deployed in practice. Let us therefore take a step back and look at the context of the SQF data.

Sources of bias in the SQF data

The major concern that has been identified in the literature for SQF data is how the data is generated in the first place. In their paper ”Residual Unfairness” Kallus and Zhou 2018 conceptualise the problem as shown in Figure 5. We define a person by their sensitive feature (A) and non-sensitive features (X). For each person in the population of interest a police officer decides whether to stop them or not. The problem arises in the decision to stop someone because it might be influenced by prejudice. In the SQF context we can imagine that the police is generally more suspicious towards PoC than white people. Or we can imagine that they are stopping people in general more frequently in high crime areas which happen to be correlated with low-income neighbourhoods which are more populated by PoC than white people sources. Based on this biased decision policy people are included in the sample ($Z = 1$) or they are not ($Z = 0$). But naturally, we can only know the outcome of a stop for the people who were stopped. Kallus and Zhou 2018 distinguish between target population and training population in such scenarios. The target population is the one on which we want to use the ADM on while the training population are the observations the biased decision policy chose to include in the sample and on which the algorithm is trained. In the SQF data we can see this form of bias by comparing the race distribution of NYC to the race distribution in the SQF data. From Figure 6 it is clear that in terms of race the SQF data does not represent the general population of the city. We can see that white people form the majority of the population in NYC, but only make up a tiny fraction of SQF stops. Black people in contrast are the third-largest ethnic group in NYC while they exceed any other group in the SQF data by far. Figure 6 shows that selection bias might be at play in the decision of stopping a suspect.

We take the borough as the non-sensitive variable that in combination with the race

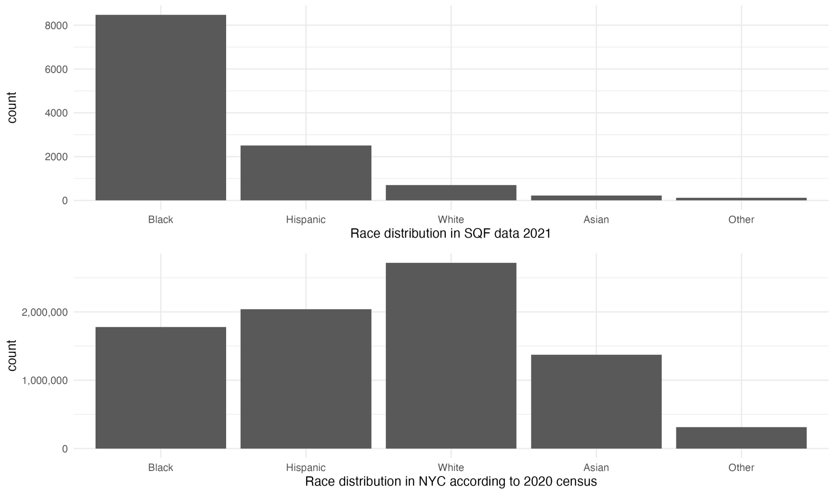

Figure 6: Comparison of race distribution in the training and target population.

describes the suspect. The comparison of the race distribution in each borough for the SQF data and the New Yorker population as a whole supports the argument that black people are stopped more leniently Figure 3. The questions now are, why the group metrics did not detect any unfairness in our algorithm and how such biases can be addressed in fairness practice?

Why the group metrics did not show any unfairness in the algorithm can be answered ins a straight forward way. Group metrics offer a rather isolated view on fairness. They assess disparities in algorithimic predictions between protected groups rather than measuring the fairness of a whole situation. Thus, group metrics are not designed to detect selection bias. They work with the joint distribution of $Y, A, \hat{Y}$ [3] and do not take any additional information into account. When we rely on the true label Y (Separation) to detect unfairness but the true label itself is not reliabel (not generated via an objective truth), then the group metrics cannot show this mechanism Castelnovo et al. 2022. To answer the second question we use the following chapter to examine two papers that propose methods to address selection bias in fairness practice and have explicity used the SQF data as a case study.

6 Residual Unfairness

We already outlined the formal problem setting in Kallus and Zhou 2018 in Figure 5. The main message of the paper supports the saying "bias in, bias out". They argue that fairness traidional fairness interventions on the training data, such as thresholding, are not enough to ensure fainress in general future applications fo the algorithm. We will take the main findings of the paper and translate them to the SQF scenario.

Additional notation: Z is the decision of the biased inclusion policy ($Z = 1$ means the subject is included into the training population). T indicated whether a person is in

[3] There are variations of group metrics that allow for non-sensitive attributes X to be considered as well when assessing fairness. An example is conditional statistical parity Verma and Rubin 2018.

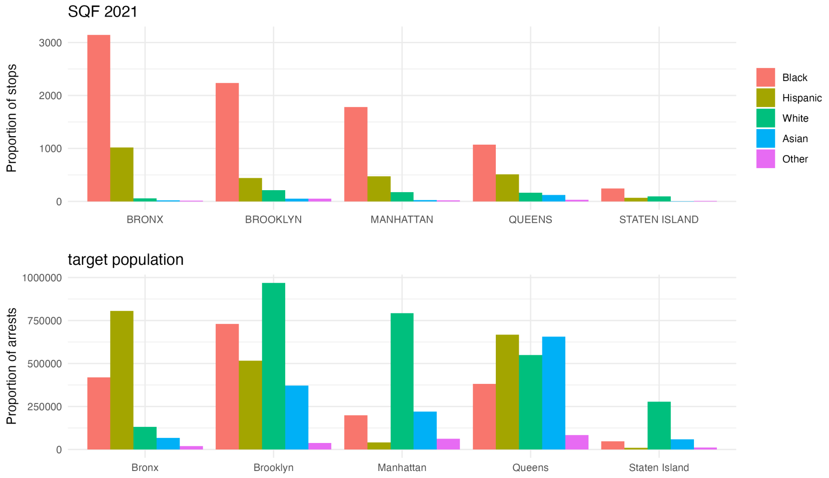

Figure 7: Distribution of race by borough in the SQF data (top) and NYC as a whole (bottom).

the target population ($T = 1$) means that we want to use the trained algorithm on the person. $\hat{R} \in [0, 1]$ is the prediction score. The problem depicted in Figure 5 can be now expressed as follows. We only have knowledge about $X, A, Y|Z = 1$ while we do not know $X, A, Y|Z = 0$. All information about the stop, the demographic details of the person and the outcome which serve as training label for an algorithm are only observed for stopped individuals.

The paper defines fairness via equal opportunity. Equal opportunity demands that the true positive rates across groups are equal, so that truly arrested individuals are predicted as such. In our case, in which a positive prediction $\hat{Y} = 1$ is undesirable it makes more sense to look at the false positive rates (or equivilantely at the true negative rates) and define fairness via predictive equality, i.e. $P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = b)$ or $P(\hat{Y} = 0|Y = 0, A = a) = P(\hat{Y} = 0|Y = 0, A = b)$. The paper looks at a thresholding classifier, if the prediction score exceeds a certain threshold the positive value for the target is predicted. This allows us express the false positive rate and the true negative rate of a group a with respect to an event E via the cummulative distibution function. $F_a^E = P(\hat{R} \leq \theta|Y = 1, A = a, E)$. Note that we condition on the truly negative subject in the sample. In the SQF case this would mean that we only look at people that were innocent, so not arrested. The truly arrested that have a predicted probability $\hat{R} \leq \theta$ are wrongly classified as not-arrested, while the ones for whom $\hat{R} > \theta$ are correctly classified as arrested. F_a^Z gives us nothing other than the false negative rate in the training population and F_a^T is the false negative rate in the target population. When we want to define a predictive equality classifier on the training population we require $F_a^Z(\theta_a) = F_b^Z(\theta_b)$ to hold.

An optimal derived predictive equality classifier can then be defined as $\hat{Y} = I(\hat{R} > \theta_A)$ and $F_a^Z(\theta_a) = F_b^Z(\theta_b)$ for all groups a,b. In words, the classifier predicts the positive (undesirable) outcome with a group-specific threshold for each member of the group while this group specific threshold is set in such a way that predictive equality on the training data is fulfilled. We will not go into detail of how one finds such a classifier but refer to

Hardt et al. 2016 who propose a post-processing method to derive the optimal thresholds for each group.
With the definition of fairness as predictive equality (equal tnr/fpr across groups) we can in turn define unfairness as nothing other than the difference in true negative rates between groups, i.e. $\epsilon_{a,b}^{E} = P(\hat{Y} = 0|Y = 0, A = a, E) - P(\hat{Y} = 0|Y = 0, A = b, E)$ and call this inequity of predictive equality. $\epsilon_{a,b}^{T=1} > 0$ shows discrimination against group b, since this means that the true negative rate for group b is lower than for group a. When we construct an equal opportunity classifier (via some fairness intervention) then $\epsilon_{a,b}^{Z=1} = 0$ holds for this classifier. This also means that any unfairness that might show in the target population cannot be explained via existing ineuqities between groups in the training population but via existing differences in the training and the target population. So it is unfairness that gets introduced when we try to generalize our algorithm to the population it was not trained on. Kallus and Zhou 2018 call this residual unfairness, since this is unfairness remaining event after fairness adjustments.

Strong disparate benefit of the doubt

To fully understand the results of the paper, we additionally introduce the concept of stochastic dominance, originating from decision theory. Let F, G be two cummulative distribution functions. Then G first order stochastically dominates $F \preceq G$ when $F(\theta) \geq G(\theta)\ \forall\theta$. Recall that G and F are cumulative distribution functions. So first order stochastic dominance of G over F, smaller values of G for each input value θ, that the population described by the CDF of G consistently has higher probability values than population F. Their probability mass is concentrated towards the higher input values thus the cummulative distribution function is small for small input values. Equipped with these definitions the paper constructs difference scenarios of (un)fairness. The bottom line is always that equal opportunity in the training population does not guarantee equal opportunity in the target population. We will introduce the one scenario mathemtically aligning the statements of the paper with the sqf scenario and will only conceptually explain the oterh scenarios which are extensions of the first one.
We assume the following: $F_a^{Z=1} \succeq F_a^{T=1} and F_b^{Z=1} \preceq F_b^{T=1}$ and at least one of the equalities does not hold (either $F_a^{Z=1} \neq F_a^{T=1}$ or $F_b^{Z=1} \neq F_b^{T=1}$ or both). Then every derived equal opportunity classifier has nonnegative inequity of predictive equality for group b relative to group a $\epsilon_{a,b}^{T=1} \geq 0$ and at least one derived equal opportunity classifier will have a strictly positive inequity of predictive equality disadvantaging group b relative to group a $\epsilon_{a,b}^{T=1} > 0$.
In words this means that for group a in the training population we have way more people with high scores (propbabilities of getting positive (undesirable) prediction) than in group a of the target population. So group a members were stopped very carefully. For group b members the opposite is true. In the train population of group b, there are many more people with low risk scores than there are in the target population of group b. This means group b members were stopped very leniently. This aligns with the fact that the sqf data records considerably more stops for black people than white people. In this case the propositions of the paper will show us again that even after adjusting for equal error rates, the classifier will disadvantage group b when applied to the target population. The results of the paper would then say that adjusting a classifier for predictive equality, so

equal false positive rates across groups, is not enough to ensure fairness on a whole and in future application.
The paper admits that the assumptions are strict and in reality unlikeliky to be met. The assumptions would mean that the police is so biased against group b members that the proportion of low-risk group b members among stopped individuals is higher than the proportion of low-risk gorup b members in the general population. This would require a very unreasonable stopping policy. Therefore in the propositions that follow they weaken the assumptions. They allow that the stochastic dominance works in the same direction for both groups but the difference in training and target population for group b is so much more different than for group a that discrimination persists. Recall that the fairness definition of the paper is based on true positive rates and they are especially interested in the post-processing method by Hardt et al. The method takes as input the error rates of a classifier (e.g. the false positive and true positve error rates) in order to find group-specific thresholds that equalise these error rates across groups. If we now put in the "wrong" error rates, the error rates of the training population, which, however, are nto representative for the target population due to selection bias, we estimate the wrond thresholds. Therefore Kallus and Zhou offer a way to estimate the error rates in the target population based on training data and some additional information. These "corrected" error rates can then be used to create fairness interventions that will also spill over to the target population. Their approach is interesting and they use it on the SQF data themselves, but it is also very tailored towards error-rate-based group metrics and is mostly useful when the fairness method relies on the error rates, such as the thresholding by Hardt et al. When the fairness methods does not take error rates as an input, there is little use in estimating them for the target population. Though it could be interesting regardless, to get picture for how the algortihm could generalise.
As the fairness audit in the previous chapter showed little disparities, we only take their method to estiamte the generalisation of the algorithm. We match the target population to the training population via precinct. .... Results: On the target population FPR for POC is estimated to be higher than for white people, but true positive rate also. FNR is estimated lower for POC than for white people on the target population and TNR is estimated to be lower for POC than for white people. So in general POC in the target population are more often expected to be more often predicted as positive and less often as negative.
For the train population the FPR of black and white people is basically the same whereas the difference in estimated FPR for the target population is a larger (still small) with POC having the higher FPR than white people. High FPR is undesirable. But at the same time the TPR in the train population is basically the same between groups whereas the for the target population POC get estimated a higher TPR than white people. So maybe it is better put this way: It is estimated that the classifier performs better in terms of TPR on the POC of the target population than on the white people of the target population. Naturally with a higher TPR comes also a higher FPR (inherent trade-off). So the estimation overall show minor differences in race groups and at most maybe even a slightly better estimated performance for POC than for white people.

Bias in, bias out - an alternative perspective

An interesting perspective on this observation can be found in Rambachan and Roth 2020. They take a different stance on the problem of biased training data than Kallus and Zhou 2018 and question the ”bias in, bias out” mechanism.

They formalise the problem as follows. For the decision-maker (the police) an individual is characterised by the random vector (X, U, A), where X and A have the same meaning as in Kallus and Zhou 2018 and U is a set of unobserved features. These latent variables are unknown to the algorithm but are characteristics the police bases their decision to stop someone on. In the SQF context this could be the personal impression the officer got of a suspect which is not recorded and hard to measure.

The paper assumes the police is a taste-based classifier against African-Americans. This means they hold some form of prejudice against the group of African-Americans that influences their decision to stop a member of this group. In general, for stopping any person, an officer incurs a cost c ¿ 0. If they stop an individual that turns out to be involved in criminal activity and is therefore arrested, the officer receives a reward b = 1 [4]. In case of stopping an innocent person b = 0. For stopping African Americans the payoff an officer expects increases by $\tau > 0$ compared to stopping a white person. The total payoff for stopping an individual is given by:

$$Y + \tau * A - c$$

where Y is the outcome of the stop, τ is the discrimination parameter, $A \in \{0, 1\}$, and $c > 0$ is the cost for stopping a person. Holding the costs c and the outcome of the stop Y constant, searching an African American results in a higher payoff than searching a white person. The goal of the police is to maximise their payoff. Therefore they stop an individual according to the following threshold rule:

$$Z(X, U, R) = 1(E[Y|X, U, A] \geq c - \tau * A)$$

This means that the threshold for stopping an African American is *lower* than for stopping a white person. Consequently, the police stops African Americans more leniently than white people. This taste-based discrimination rule is the biased decision policy introduced in Kallus and Zhou 2018. In Rambachan and Roth 2020 the authors speak of ”selective labels” where again the tupel (Y, X, A, Z) is only available for $Z = 1$.

In short: It actually depends on the outcome and the training sample whether the discrminiation of the previously discriminated (bias interhitance) exists. In some cases it can actuall come to the opposite effect, which they call bias reversal. The mechanism is as follows: the historically discriminated groups is very represented in the sample as being included is an act of discrimination itself. This means we have more training data for the disadvantaged group, they resemble the target population more, as they were more leniently included, and thus the classifier generalised better to the disadvanteged group. When we collect more data for the group, we come closer to the target population and our classifier will work better on the target population for the group with more data.

[4] The reward can set to any number b ¿ 0. We assume b = 1 as in Rambachan and Roth 2020 without loss of generality.

Black people are more leniently stopped, leading to higher stopping rates in for black people in the training data, meaning more training data for this group. Because we stop black peopel more leniently, we record many innocent black people in our data. In Kallus and Zhou 2018 this would lead to a lower learned threshold [5] for black individuals. Applied on the target population this would mean that we would predict too many false positive. The threshold estimated from the training data is so low that we classify to many people as guilty because in the target populations the scores are actually higher and meet the threshold easily. In Rambachan and Roth 2020 they say that by stopping (searching, they actually talk about searching, not stopping) black people so leniently, our sample for black people comes actually pretty close to the target population. In other words, the training data for black people is pretty close to the target data for black people, which means that our classifier will work well on the target population for black people.

To summarise, in Kallus and Zhou 2018 bias against a group results in a less representative sample. In Rambachan and Roth 2020 bias against a group results in a more representative sample.

Theorem 1

The prediction for african americans is weakly decreasing in tau. This means, as tau increases (so racial bias increases), the expected value for Y gets actually lower, so closer to zero, so less often predicted to have a contraband. What is happening? Higher tau means lower searching threshold for african americans. So the data for african americans becomes "more noisy", more and more innocent people come into our sample, so we predict lower risk for african americans. In Rambachan and Roth 2020 paper this translates to a more representative training data for african americans and thus also better performance on the general population of african americans. In Kallus and Zhou 2018 paper the mechanisms is the same, we also estimate lower risks cores for african americans, but then sth else happens. I think in Kallus we then do a fairness intervention that leads us to setting a LOWER threshold for african americans, meaning we predict them as guilty more easily to achieve the same FPR as in the other group. I think in kallus they first formulate it in the strict way, where the police is so biased against african americans that the stopped african americans are LESS likely to actually have a weapon than the general population. But they relax this setting afterwards.

What happens if we train the logistic classifier (to predict weapon yes no) on the SQF as is (Kallus), do not do a post processing fairness intervention (NO Hardt et. al) and test the classifier on the target population (that is created via the weighing method of Kallus and Zhou)? I think according to Rambachan and Roth 2020 we should observe bias reversal.

[5] first this leads to lower risk scores for black individuals. And then via fairness adjustments (e.g. for equalized odds) this leads to lower thresholds for black individuals.

List of Figures

List of Tables

Acknowledgement

A Electronic Appendix

Data, code and illustrations are available in electronic form.

References

Castelnovo, Alessandro et al. (Mar. 2022). "A Clarification of the Nuances in the Fairness Metrics Landscape". In: *Scientific Reports* 12.1, p. 4209. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07939-1. (Visited on 12/23/2024).

Caton, Simon and Christian Haas (July 2024). "Fairness in Machine Learning: A Survey". In: *ACM Computing Surveys* 56.7, pp. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3616865. (Visited on 12/23/2024).

Fernando, Martínez-Plumed et al. (2021). "Missing the Missing Values: The Ugly Duckling of Fairness in Machine Learning". In: *International Journal of Intelligent Systems* 36.7, pp. 3217–3258. ISSN: 1098-111X. DOI: 10.1002/int.22415. (Visited on 12/10/2024).

Hardt, Moritz et al. (2016). "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc. (Visited on 01/27/2025).

Kallus, Nathan and Angela Zhou (July 2018). "Residual Unfairness in Fair Machine Learning from Prejudiced Data". In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, pp. 2439–2448. (Visited on 12/24/2024).

Kusner, Matt J et al. (2017). "Counterfactual Fairness". In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA. URL: https://papers.nips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Rambachan, Ashesh and Jonathan Roth (2020). *Bias In, Bias Out? Evaluating the Folk Wisdom.*

Verma, Sahil and Julia Rubin (May 2018). "Fairness Definitions Explained". In: *Proceedings of the International Workshop on Software Fairness*. Gothenburg Sweden: ACM, pp. 1–7. ISBN: 978-1-4503-5746-3. DOI: 10.1145/3194770.3194776. (Visited on 11/16/2024).

Zafar, Muhammad Bilal et al. (2017). "From Parity to Preference-based Notions of Fairness in Classification". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. (Visited on 12/29/2024).

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, February, 24th 2025

Juliet Fleischer