

Seminar Thesis

---

# FairML and the SQF dataset

---

Department of Statistics  
Ludwig-Maximilians-Universität München

**Juliet Fleischer**

Munich, February, 22th 2025



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

## **Abstract**

In this study we provide an introduction to the most common fairness definitions and subtleties that come with them. We advocate for tackling fairness in a wholeistic way, taking into account how the data was generated and how it will be used. This will be illustrated by a case study on the Stop, Question, and Frisk data (SQF) from the New York Police Department (NYPD).

## Acknowledgement

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Fairness Metrics</b>	<b>1</b>
2.0.1	Separation . . . . .	2
2.0.2	Sufficiency . . . . .	3
<b>3</b>	<b>Fairness Methods</b>	<b>6</b>
<b>4</b>	<b>Case Study: Stop, Question, and Frisk</b>	<b>7</b>
<b>5</b>	<b>Residual Unfairness</b>	<b>12</b>
<b>A</b>	<b>Electronic Appendix</b>	<b>V</b>

## List of Figures

1	Density of predicted probabilities both groups. . . . .	8
2	Comparison of fairness metrics. . . . .	9
3	Comparison of arrestment rates for 2023 (left) and 2011 (right). . . . .	10
4	The bias loop. . . . .	10
5	Selection bias in the SQF data. . . . .	11
6	Comparison of race distribution in the training and target population. . . .	12

## List of Tables

# 1 Introduction

## 2 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. We can distinguish

- 1) group fairness vs. individual fairness
- 2) observational vs. causality-based criteria Castelnovo et al. 2022

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 the data to assess fairness while causality-based criteria make assumptions about the causal structure of the data and base their notion of fairness on this (so basically observational says, fairness is when I can measure equality from my distribution and causality says fairness is when the cause for my decision is not discriminatory against someone or a group). On the basis of these fundamental ideas, a plethora of formalisations 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. Specifically, we want to use the following running example, inspired by our case study on real data in chapter x. The crime rates in NYC should be decreased with the help of a new AI tool. Specifically, the administration orders a team of machine learning experts to design an automated decision-making system that should predict criminal activity of a person. It should be employed by police officers to decide whether to stop a person or investigate them further. Past police stops serve as training data. Given the history of racial profiling in the United States, it is reasonable to raise concerns about racial decision patterns the algorithm could learn from. First, we approach this from a group fairness perspective.

### Group fairness

The notion of fairness underlying group metrics is that discrimination of certain groups of the population defined via the protected attribute should be prevented. Group fairness can be grouped into three main categories, independence, separation, and sufficiency.

#### 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  $\hat{Y} \perp A$ . 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 formalised as demographic parity/statistical parity  $P(\hat{Y}|A = a) = P(\hat{Y}|A = b)$ . The other two groups of group fairness metrics, Separation and Sufficiency can both be derived from the error matrix.

### 2.0.1 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  $P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b)$  mlr3: `fairness.fnr`, `fairness.tpr`
- Predictive equality/ False positive error rate balance:  $P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = b)$  or  $P(\hat{Y} = 0|Y = 0, A = a) = P(\hat{Y} = 0|Y = 0, A = b)$  mlr3: `fairness.fpr`, `fairness.tnr`
- Equalized odds:  $P(\hat{Y} = 1|Y = y, A = a) = P(\hat{Y} = 1|Y = y, A = b) \forall y \in \{0, 1\}$  mlr3: `fairness.equalized.odds`
- Overall accuracy equality:  $P(\hat{Y} = Y|A = a) = P(\hat{Y} = Y|A = b)$  mlr3: `fairness.acc`
- Treatment equality:  $\frac{FN}{FP}|_{A=a} = \frac{FN}{FP}|_{A=b}$

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. fulfilled, one could also define equal opportunity via the true positive rate. 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. With its focus on the false positive rates, predictive equality is also presented in the context of punitive tasks. Since people could experience potential harm on the basis of a positive prediction, the proportion of truly innocent people that do not deserve punishment should be kept at a minimum. For assistive tasks, such as deciding who receives some kind of welfare, a focus on minimising the false negative rate could be more relevant. 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. Treatment Equality is another variation that forms the error ratio for each group and requires it to be equal. Finally, overall accuracy equality simply requires equal accuracy between groups, meaning equal proportion of correctly classified individuals in each group.



## 2.0.2 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. This leads to the following fairness metrics:

- Predictive parity/ outcome test:  $P(Y = 1 | \hat{Y} = 1, A = a) = P(Y = 1 | \hat{Y} = 1, A = b)$   
mlr3: `fairness.ppv`
- Equal true negative rate:  $P(Y = 0 | \hat{Y} = 0, A = a) = P(Y = 0 | \hat{Y} = 0, A = b)$  mlr3: `fairness.npv`
- Equal false omission rate:  $P(Y = 1 | \hat{Y} = 0, A = a) = P(Y = 1 | \hat{Y} = 0, A = b)$  mlr3: `fairness.fomr`
- Equal false discovery rate:  $P(Y = 0 | \hat{Y} = 1, A = a) = P(Y = 0 | \hat{Y} = 1, A = b)$
- Conditional use accuracy equality:  $P(Y = 1 | \hat{Y} = 1, A = a) = P(Y = 1 | \hat{Y} = 1, A = b) \wedge P(Y = 0 | \hat{Y} = 0, A = a) = P(Y = 0 | \hat{Y} = 0, A = b)$

Predictive parity requires that the probability of actually being positive, given a positive prediction is the same between groups. Following the same principle, we can require that the probability of actually being negative, given a negative prediction is the same between groups. If we instead look at errors again, we can require equal false omission rates between groups or equal false discovery rates between groups. 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. By not only requiring one of these criteria but two simultaneously, we can build a stronger metric, like conditional use accuracy equality that requires same positive predictive values between groups and same negative predictive values between groups. 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 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.

- Calibration:  $P(Y = 1 | S = s, A = a) = P(Y = 1 | S = s, A = b)$
- Well-calibration:  $P(Y = 1 | S = s, A = a) = P(Y = 1 | S = s, A = b) = s$
- Balance for positive class:  $E(S | Y = 1, A = a) = E(S | Y = 1, A = b)$
- Balance for negative class:  $E(S | Y = 0, A = a) = E(S | Y = 0, A = b)$

Calibration requires that the probability for actually being positive, given a score  $s$  is the same between groups. 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. Well-calibration is a stronger version of this, requiring that the probability for actually being positive, given a score  $s$  is the same between groups and equal to the score itself. This means, when for a set of suspects the classifier predicts a certain probability  $s$  of crime, then the proportion of people that actually committed crime should be  $s$ . Balance for the positive class takes the expectation over the predictions scores of the people that are actually positive and wants them to be equal across groups. We do not want that one of the positive people of one group get on average a higher score than the positive people of another group. The same holds for the negative class, formalized as balance for the negative class. To contrast 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 this enforcement has the intended benefits, especially over the long term.

reference?

## Individual fairness

If we want to equalise 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 it 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. Individual metrics therefore shift the focus. The underlying idea of fairness is that similar individuals should be treated in a similar way. Different individuals should be treated in a different way. It is an intuitive idea that was already formulated by Greek philosopher [bothmann citation](#). **Fairness through awareness (FTA)** formalizes this idea as Lipschitz criterion.

$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$

This simply 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 similar in their features, they should also get similar predictions from the algorithm. The challenge of FTA is the definition of precisely this equality in the feature space. Defining when two individuals are similar is not much different from defining fairness in the first place Castelnovo et al. 2022. There is no clear solution to this. In any case, the choice of  $d_X$  should take context-specific information into account. We want to find a distance metric, that suits the target and represents an ethical formalisation of similarity in the features.

**Fairness through unawareness (FTU) or blinding** This is primarily formulated as

procedural rule. Blinding tells us to not use the protected attribute explicitly in the decision-making process. So at first this would simply mean to discard the protected attribute from the data during training. After training, FTU can be tested by simulating a doppelganger for each person in the dataset. This doppelganger has the exact same features with the exception of the protected attribute, which is flipped (easy in binary PA case). If both these instances have the same prediction, the algorithm would satisfy FTU Verma and Rubin 2018. This is actually also a form 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. Blinding is however to be seen critically. FTU has the problem of proxies. These are variables that are strongly correlated with the protected attribute. Therefore, it's not enough to simply mask the information of the sensitive attribute during training because discrimination can persist via these proxies. This becomes clearer, when imagine that we remove information, such that this feature is simply not available to the classifier during training. The place of residence, however, is strongly correlated with the person's ethnicity. Thus, indirect discrimination based on ethnicity remains, even though the information was not directly available during training. Suppression therefore extends the idea of blinding and the goal is to develop a model that is blind to the sensitive attribute and the proxies. The drawback is, that it is unclear when a feature is sufficiently correlated with the sensitive attribute to be counted as proxy. Additionally, we could lose important information by removing too many these features Castelnovo et al. 2022. 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. For this paper we only mention the group metrics to give the reader a more complete picture of the fairness metric landscape. For more information we refer to Kusner et al. n.d.

**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. n.d.) The two most common individual fairness metrics are counterfactual fairness and path-based fairness.

## 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 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 a best one depending on the

context and the data. The next section will present ways to mitigate algorithmic bias once detected by one of the fairness metrics.

## 3 Fairness Methods

### Fairness methods

Another question fair machine learning deals with is how algorithms can be adjusted such that they fulfill 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. Reweighting 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 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 fulfilled. 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.

## 4 Case Study: Stop, Question, and Frisk

### Stop, Question, and Frisk data

The legal sector is one in which ADMs 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 criticised 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 criticised for over-targeting 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. 34 variables remain that provide us with information about the stop and demographic information of the stopped person. From this reduced dataset we filter out the complete cases and end up with 12039 observations.<sup>1</sup> We choose the arrestment of a suspect as target and the race as protected attribute. For the fairness audit later in the chapter we dichotomise the PA to adjust our situation to the common binary classification, binary PA scenario in the fairness literature. For the descriptive analysis we leave align the race description of SQF data with the 2021 census data. Thereby placing "Black Hispanic" into the group "Black" and summarising "American Indian/ Native American" of SQF and "Middle Eastern/ Southwest Asian" of SQF 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

<sup>1</sup>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.

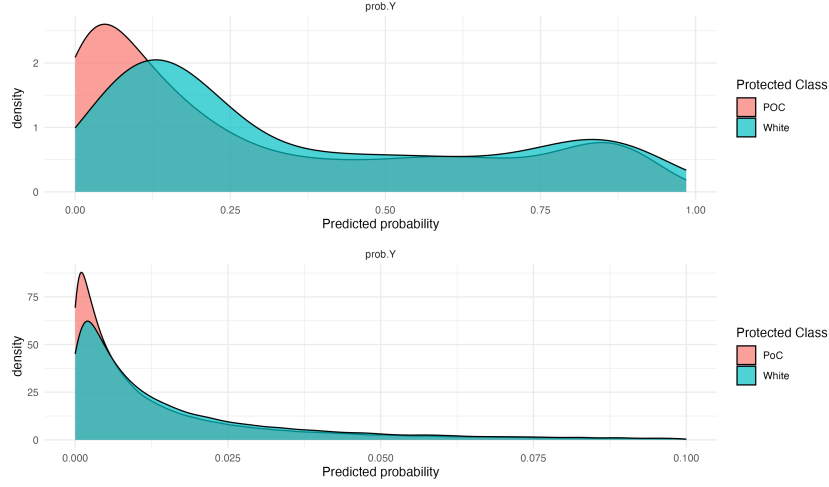


Figure 1: Density of predicted probabilities both groups.

fairness metrics are observational and constructed from the joint probability of  $Y, \hat{Y}, A$  alone, this already gives us a hint that the classifier will show little racial disparities.

## Fairness Auditing

To train a random forest classifier on the 2023 data to predict the arrest of a person, we dichotomise the race attribute by grouping "Black" and "Hispanic" as people of colour ("PoC") and "White", "Asian", and "Other" as white ("White"). 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  $\epsilon$  in practice. Especially the error rates (`fmr`, `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 visualise 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 tends 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 equalised 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  $\epsilon$  to choose. Given that the classifier is fair from a group perspective, it does not make 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.

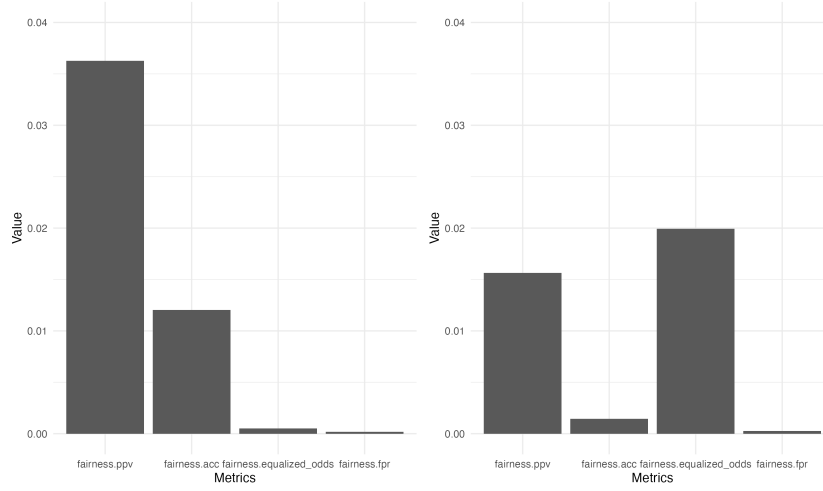


Figure 2: Comparison of fairness metrics.

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 rates 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 cutoff 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, equalised odds has the greatest 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 this data to make an optimal prediction, on which the decision-makers base their judge-

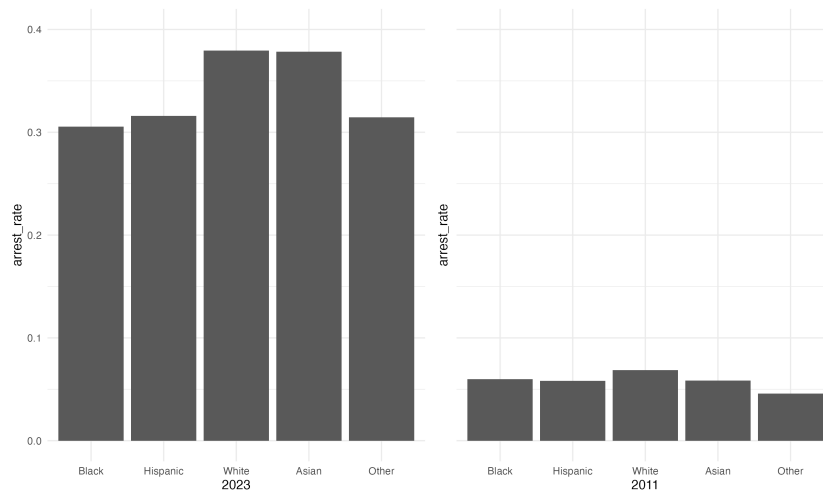


Figure 3: Comparison of arrestment rates for 2023 (left) and 2011 (right).

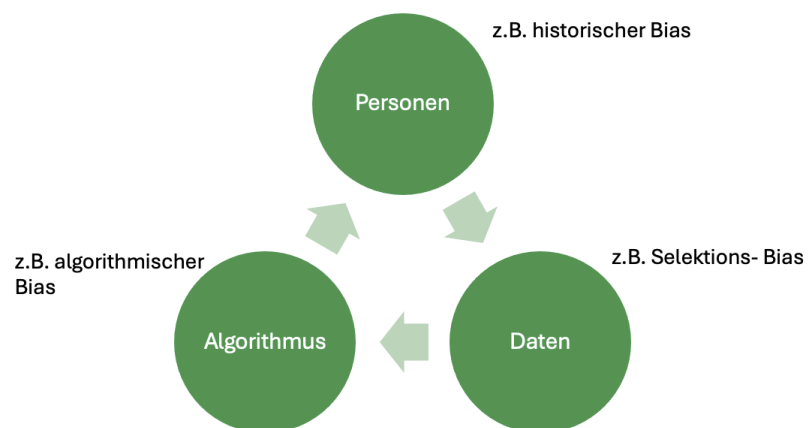


Figure 4: The bias loop.



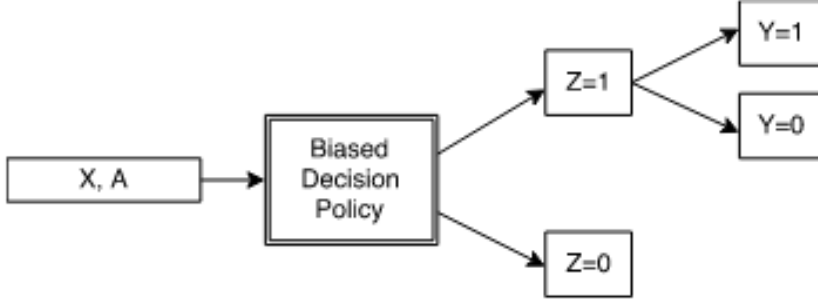


Figure 5: Selection bias in the SQF data.

ment 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 n.d. 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. 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, which constitutes the target of a classification task, for the people who were stopped. Kallus and Zhou n.d. 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.

To get a more nuanced picture we also plot the true arrestment rate per ethnic group in the SQF 2023 data and find that White and Asian people have the highest arrestment rates and black people the lowest. This supports the argument that black people are stopped more leniently and there is a biased decision policy in place. 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 in a straight forward way. Group metrics offer a rather

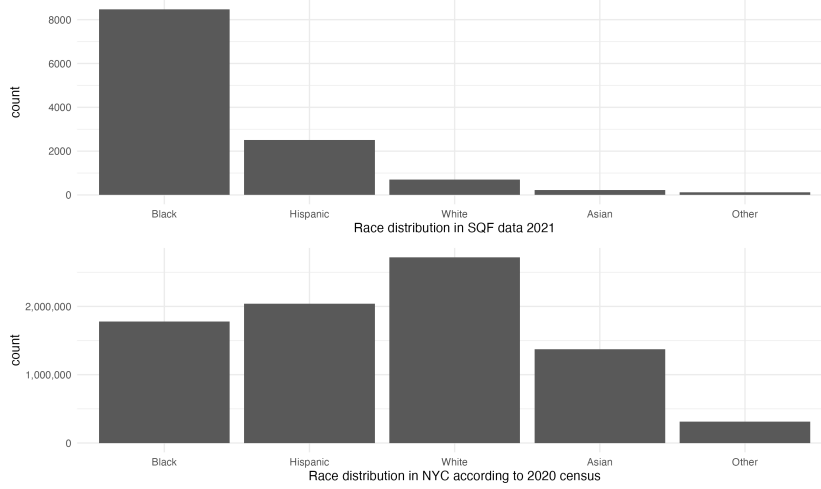


Figure 6: Comparison of race distribution in the training and target population.

isolated view on fairness. They assess disparities in algorithmic 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  $Y, A, X, \hat{Y}$  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 reliable (generated via an objective truth), then the group metrics cannot show this 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 explicitly used the SQF data as a case study.

## 5 Residual Unfairness

We already outlined the formal problem setting in Kallus and Zhou n.d. in Figure 5. The main message of the paper supports the saying "bias in, bias out". They argue that fairness traditional fairness interventions on the training data, such as thresholding, are not enough to ensure fairness in general future applications for 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 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 equivalently 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 cumulative distribution 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 inequities 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 n.d. 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 cumulative 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  $\theta$ , 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 cumulative 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 mathematically aligning the statements of the paper with the sqf scenario and will only conceptually explain the other 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 (probabilities 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 unlikely 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 group 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 positive 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 not representative for the target population due to selection bias, we estimate the wrong 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 algorithm could generalise.

As the fairness audit in the previous chapter showed little disparities, we only take their

method to estimate the generalisation of the algorithm. We match the target population to the training population via precinct. .... Result: the classifier has even lower estimated error rates for the target population than for the training population - what is happening here?

## Bias in, bias out - an alternative perspective

A potential answer can be found in Rambachan and Roth n.d. They take a different perspective on the problem of biased training data than Kallus and Zhou n.d. and question the "bias in, bias out" mechanism.

In short: It actually depends on the outcome and the training sample whether the discrimination of the previously discriminated (bias inheritance) exists. In some cases it can actually 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 disadvantaged 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.

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 people more leniently, we record many innocent black people in our data. In Kallus and Zhou n.d. this would lead to a lower learned threshold<sup>2</sup> 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 too many people as guilty because in the target populations the scores are actually higher and meet the threshold easily. In Rambachan and Roth n.d. 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 n.d. bias against a group results in a less representative sample. In Rambachan and Roth n.d. 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 n.d. paper this translates to a more representative training data for african americans and thus also better performance on the

---

<sup>2</sup>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.

general population of african americans. In Kallus and Zhou n.d. 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), don't 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 n.d. we should observe bias reversal.

## 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 (n.d.). “Residual Unfairness in Fair Machine Learning from Prejudiced Data”. In: ().
- Kusner, Matt J et al. (n.d.). “Counterfactual Fairness”. In: ().
- Rambachan, Ashesh and Jonathan Roth (n.d.). “Bias In, Bias Out? Evaluating the Folk Wisdom”. In: ().
- 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, 22th 2025

---

Juliet Fleischer