Seminar Thesis

FairML and the SQF Dataset

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, February, 28th 2025

Submitted in fulfilment of the requirements for the degree of B. Sc.
Supervised by Dr. Ludwig Bothmann

Abstract

With the growing presence of AI in our society, issues of social justice and fairness are increasingly intersecting with technical research fields. 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 analysis leads us to examine the problem of selection bias and its impact on algorithmic learning. To further explore this issue, we review studies that have worked with the SQF dataset and uncovered key theoretical findings, such as residual unfairness and bias reversal. The main contribution of this thesis lies in comparing and contrasting the various approaches used to study fairness in the context of the SQF dataset. Rather than identifying a single ”correct” method, we emphasize the importance of understanding the reasoning, assumptions, and implications behind different fairness frameworks.

Contents

1 Introduction

The challenge of creating a fair and equitable society has concerned people since ancient times. With the rise of artificial intelligence (AI) questions of justice and fairness have taken on new urgency. AI enables automated decision-making systems (ADM) that are now common in law, healthcare, finance, and other fields, where they can affect the lives of individuals significantly. Despite their ongoing improvements they carry the risk of perpetuating and even exacerbating social injustices.

After a general introduction to the study of fairness in machine learning (fairML), this paper turns its focus to the Stop, Question, and Frisk (SQF) dataset published yearly by the New York Police Department (NYPD). Since 1990 the US Supreme Court has been allowing police officers in New York City to stop individuals if they have a reasonable suspicion that they are involved in criminal activity (Terry v. Ohio (1968), 392 U.S. 1, U.S. Supreme Court).

While proponents argue that the stop-and-frisk strategy is an effective crime prevention tool, many criticize the police for disproportionally targetting people of colour (PoC). In 2013, a federal district court ruled that the way stop-and-frisk was implemented in NYC between 2004 and 2012 was unconstitutional, violating both the Fourth and Fourteenth Amendments.[1] Official statistics show the steep decline in stops following the 2013 judgement. They have remained at a low level ever since.[2]

The main contribution of this thesis lies in reviewing multiple studies that examine the fairness of SQF from different angles. Though these studies seek to answer the same question—"Is the Stop, Question, and Frisk practice fair?"—they approach the problem differently and arrive at alternative conclusions.

This divergence is not necessarily a contradiction but rather a reflection of the diverse perspectives and objectives that shape fairness research. Each study addresses fairness within its own problem setting, making its conclusions valid within that specific context. However, this can create confusion, as studies with different assumptions and goals may still claim to answer the same overarching question. Our objective is not to identify *the right* approach but to emphasize the importance of understanding data context and problem framing when evaluating fairness.

The paper is organized in the following way: in Section 3, we introduce the most common fairness metrics and techniques used in machine learning. Next, in Section 4 we apply the theoretical concepts to the real-world SQF dataset. The application on real-world data will show difficulties that come with fairness in practice. This will lead us to explore other studies that have worked with SQF data in Section 5.

[1] Link to Report

[2] https://www.nyclu.org/data/stop-and-frisk-data

2 Related Work

Fairness in machine learning has attracted considerable attention in recent years, leading to a rich literature of definitions and evaluation frameworks. Several works provide broad overviews of these definitions. For example, Verma and Rubin 2018 offer a comprehensive overview of the most used fairness metrics, accompanied by a case study on the German Credit Dataset. Castelnovo et al. 2022 build on this work and highlight the nuances and subtleties that come with typical fairness metrics. The work of Corbett-Davies et al. 2023 and of Barocas, Hardt, and Narayanan 2023 serves as detailed resources that offer deeper insights into common fallacies in fairML.

Beside the definition of fairness a major branch of research has concerned itself with the design of bias migitation techniques. Mehrabi et al. 2022 and Caton and Haas 2024 provide a detailed review. Additionally, the chapter on algorithmic fairness in the `mlr3book` by Pfisterer 2024 serves as an accessible introduction to the practical implementation of fairness metrics and methods.

Beyond these more general works, a number of studies from the fields fairML, Statistics, and Economics, have focused on the SQF dataset. Goel, Rao, and Shroff 2016 use advanced statistical methods to support the claim that non-white individuals are disproportionately targeted by the New Yorker police. Khademi et al. 2019 examined fairness in SQF from a causal perspective. Their study supports the complexity of this topic as they arrive at divergent conclusions, depending on which of their metrics they use.

In the course of this paper it will become clearer that selection bias is a major concern for the SQF data. The general effects of selection bias on fairness and potential ways to counteract them have been studied by Lakkaraju et al. 2017 and Favier et al. 2023. The other studies that explicitly use SQF data, namely Badr and Sharma 2022; Rambachan and Roth 2016; Kallus and Zhou 2018 will be more closely examined in Section 5.

3 Fairness Metrics and Methods

It is easy to get overwhelmed by the sheer amount of fairness definitions in machine learning. This chapter groups the metrics in an intuitive way and motivates them in the hope to bring some clarity to the readers. What all the metrics have in common is that they build on the idea of a protected attribute (PA) or alternatively called sensitive attribute. This is a feature present in the training data because of which individuals should not experience discrimination. Examples for sensitive attributes are race, sex and age. Castelnovo et al. 2022 suggest that fairness metrics can be categorized along two essential axes:

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

Loosely speaking, group fairness aims to create equality between groups and individual fairness aims to create equality between two individuals within a group. Group membership is encoded by the PA. Observational fairness metrics provide a descriptive fairness assessment by relying on the realized distribution of random variables characterizing the population of interest. Causality-based notions, on the other hand, define fairness based on assumptions about the underlying causal structure of the data (Castelnovo et al. 2022). On the basis of these fundamental ideas, a plethora of formalizations has emerged. Most of them concern themselves with defining fairness for a binary classification task and one, often dichotomized, sensitive attribute. For this work, we will also stay within this setting. Moreover, our focus will lie on the observational metrics, as causal notions of fairness require more involved techniques that are out of the scope of this paper.
Throughout this thesis, we denote the categorical sensitive attribute as $A \in \{a, b\}$ while we assume for simplicity that it is binary. The remaining features are encoded as $X \in \mathcal{X}$. We define $f : \mathcal{X} \times \{0, 1\} \to [0, 1]$ as a prediction function that returns a score $s = f(x, a)$ representing the estimated probability that the true label is the positive class (1) for each input (x, a).
To obtain a hard prediction from the score, we define a thresholding function

$$g(s) = \begin{cases} 1, & \text{if } s \geq c, \\ 0, & \text{if } s < c, \end{cases}$$

where $c \in [0, 1]$ is a predetermined threshold (often $c = 0.5$). Thus, the final predicted label is given by $\hat{y} = g\big(f(x, a)\big) = \mathbf{1}\{f(x, a) \geq c\}$.
To facilitate the understanding of the following fairness metrics, we already frame them in the context of SQF. We define the sensitive attribute as $A \in \{\text{non-white}, \text{white}\}$. The feature set X includes all other recorded variables related to the stop, such as the time, location, and officer-related information. The true label is given by $Y \in \{\text{arrested}, \text{not arrested}\}$, while $S \in [0, 1]$ represents the predicted probability of an arrest. This probability is estimated via a prediction function f, as defined previously.

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

Table 1: Group fairness metrics

3.1 Group fairness

They observational group metrics presented in this section can be separated into the three main categories shown in Table 1, depending on which information they use.

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. This is fulfilled when for each group the same proportion is classified as positive by the algorithm. For a binary classification task with binary sensitive attribute this can be formalized as

- *Demographic parity* requires equal positive prediction ratios (ppr) for both groups:

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

Demographic parity can be appropriate, when a form of equality between groups should be enforced, regardless of context or any potential personal merit. This can be seen as fairness *in principle*, which becomes relevant when we come to the normative conclusion that the status quo should be changed. In other situations demographic parity might seem everything but fair. Think for example of a situation in which the police stops African-Americans very leniently, leading too many innocent African-Americans in the sample, while they stop white people more strictly. In this case the arrestment rate for white people should be higher than for non-white people, which would violate demographic parity. Enforcing it, however, would mean that more innocent African-American would wrongly be predicted as arrested.
In many cases it can make sense to allow for additional information to be taken into account. In our running example this could mean to require independence of the decision on race only for PoC and white people within the same borough. Like this we could account for location-specific crime rates. Therefore, an extension of demographic parity can be defined as

- *Conditional statistical parity* requires:

$$P(\hat{Y} = 1 \mid E = e, A = a) = P(\hat{Y} = 1 \mid E = e, A = b)$$

E is a set of legitimate features that encapsulates valuable information about the target Y. As mentioned, we could set $(E = borough)$ to account for the fact that certain areas of the city have higher crime rates and thus higher arrestment rates are legitimate. Notice that so far, we have only worked with $(\hat{Y}, A, X)$. The other two categories of group fairness metrics additionally make use of the true label Y and can be derived from the confusion matrix seen in Table 2.

Separation

Separation requires independence between $\hat{Y}$ and A conditioned on the true label 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* requires the false negative rates, the ratio of actual positive people that were wrongly predicted as negative, is equal between groups:

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

- *Predictive equality/ False positive error rate balance* follows the same principle as equal opportunity but for the false positives:

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

- *Equalized odds* requires that both the true positive rate and the false positive rate are equal across groups:

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

- *Overall accuracy equality* requires equal accuracy for both groups:

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

- *Treatment equality* builds groups-wise ratios of error-rates and requires equality:

$$\frac{\text{FN}}{\text{FP}}\Big|_{A=a} = \frac{\text{FN}}{\text{FP}}\Big|_{A=b}$$

The idea behind *Separation* metrics is that equality across groups does not have to hold in general but for people with the same value of the true label Y. In our running example, we require equality between PoC and white people among individuals that were (not) arrested. This means disparities across groups are allowed as long as they can be fully explained by the true label Y. Hence, Castelnovo et al. 2022 consider *separation* criteria to be most suitable when the true label Y is free from any bias, meaning it was generated via an objectively true process.

Sufficiency

Sufficiency switches the role of the true label and the prediction and requires independence between Y and A conditioned on $\hat{Y}$. Intuitively this means that we want a prediction to be equally credible between groups. When white person receive a positive prediction, the probability that it is correct should be they same as for black individuals. 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:

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

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

Table 2: Confusion matrix

- *Equal true negative rate* is based on the same principle as predictive parity. It requires that the probability of actually being negative, given a negative prediction, is the same between groups:

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

- *Equal false omission rates* requires that the probability for predicted negatives to actually have a positive label is equal across groups:

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

- *Equal false discovery rates* instead asks that the probability for predicted positives to in reality be negative is the same across groups:

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

- *Conditional use accuracy equality* combines two metrics and demands that both of the following hold:

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

Sufficiency criteria capture the decision-maker's perspective by assuming that only the prediction is available at the time of decision. For instance, police officers cannot predict the outcome of a stop when they choose to investigate someone.

While it is easy to get lost by the amount of fairness definitions in the beginning, 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 of Verma and Rubin 2018 or Castelnovo et al. 2022 but are implemented in the `mlr3fairness` package, and evidently follow the same pattern as the other metrics.

In essence the group metrics outlined so far do nothing other than picking a performance metrics from the confusion matrix and requiring it to be equal across two (or more) groups. This means that they come with trade-offs just as the usual performance metrics for classifiers do (Kleinberg, Mullainathan, and Raghavan 2017). Researchers have shown that if base rates, i.e. the proportions of the positive outcomes of the groups in the population, differ between groups, it is mathematically impossible to equalize all desirable metrics simultaneously (Chouldechova 2016). This is also referred to as the *Impossibility Theorem*.

Score-based fairness metrics

Most (binary) classifiers work with predictions scores $S \in [0,1]]$ and a hard label classifier is applied only afterwards in form of a thresholding function g. It should therefore come as no surprise that instead of formulating fairness with $\hat{Y}$ there exist fairness metrics that use the score S instead.

- *Calibration* exchanging the hard label with the score in sufficiency metrics:

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

As the score can usually take values from the whole real number line, this can in practice be implemented by binning the scores (see Verma and Rubin 2018 for an example). An extention of this idea is well-calibration:

- *Well-calibration* requires that the group wise probabilities to be equal to the score itself:

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

Choosing the right group metric

Due to the abundance of group metrics alone there are resources to assist practitioners in choosing the right metric for their specific task. One possibility is to distinguish between punitive and assistive tasks Ghani et al. 2025. For punitive tasks metrics that focus on false positives, such as predictive equality are more relevant. For assistive tasks, such as deciding who receives a welfare, the focus on minimizing the false negative rate could be more important. This points to equal opportunity as suitable metric. 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. There is dedicated work that assists in finding the right group fairness metric for a given situation and refer to Makhlouf, Zhioua, and Palamidessi 2021 for an in-depth analysis.

3.2 Individual fairness

Individual metrics shift the focus from comparison *between* groups to comparison *within* groups. 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 and is regulated by

λ. For the prediction space one could choose $d_Y(\hat{y_i}, \hat{y_j}) = |\hat{y_i} - \hat{y_j}|$ to assign same different predictions a distance of 1 and same ones a distance of 0. Less obvious is the choice of distance metric d_X in the feature space. In fact, "[...] defining a suitable distance metric $[d_X]$ on feature space embodying the concept of similarity on "ethical" grounds alone is almost as tricky as defining fairness in the first place" (Castelnovo et al. 2022).
Individuals in the SQF context could, for example, be considered similar if they live in the same borough. This would form a conceptual parallel to statistical parity, introduced in Section 3.1, where the reasonable set of feature(s) E was chosen to be the borough. This is one possible way to define similar individuals. Yet one could easily argue that taking more information, such as the criminal history and yearly income, into account is important as well.
After the choice of relevant features has been made, the next challenge is to choose the exact function d_X that appropriately captures the conceptual definition of similarity defined via the selected features. This requires domain knowledge and the key is to take context-specific information into account.
In practice, the criterion is evaluated for each individual, but in the end, we still have to build summary statistics to capture the situation. One approach is to bin the distances in the feature space. For each distance category, the average distance in the prediction space can be calculated to quantify potential disparities (see Verma and Rubin 2018 for a concrete example).

Fairness through unawareness (FTU) or blinding

In contrast to FTA, the goal of FTU is to give a direct, context-independent rule. Blinding tells us to not use the sensitive attribute explicitly in the decision-making process. Since FTU is more of a procedural rule than a mathematical definition, there exist multiple ways to test whether classifier is indeed *blind* to the PA.
One method 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 receive the same prediction by the classifier for all pairs in the data, the algorithm would satisfy FTU. In this case a parallel to a version of FTA can be seen, in which d_X is chosen to be zero only if two people are the same on all their features except for the protected attribute. Other ways to assess FTU can be found in Verma and Rubin 2018.
One limitation of FTU is that it overlooks the possible interdependence between A and X. Certain non-sensitive features can have a strong correlation with the sensitive attribute. In the presence of such variables, masking only the sensitive attribute during training is insufficient, as discrimination can persist through these proxies.
For SQF this could mean that the race attribute is removed during training, but a person's place of residence, contains information about their ethnic background. As a result, indirect racial discrimination remains, even though the information was not directly available during training. **Suppression** extends the idea of FTU and the goal is to develop a model that is blind to not only the sensitive attribute but also to its 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, important information could be lost by removing too many features (Castelnovo et al. 2022).

Comparison and Summary

Experts debate the (apparent) conflict between group and individuals fairness (see for example Binns 2020). It is out of the scope of this paper to discuss their incompatibility, and we solely point out that the sharp line we drew between group and individuals metrics gets softer as a group metric like demographic parity does not only take information from $Y, \hat{Y}, A$ into account but allows for information contained in the non-sensitive features to seep into the fairness assessment (Castelnovo et al. 2022).
Group metrics are certainly easier to understand and apply as most of them are implemented in fairness software packages. For this reason our case study in Section 4 also makes use of them.

3.3 Fairness methods

After defining fairness in a mathematical sense, the question arises how a classifier can be modified to satisfy the chosen definition. This is what fairness methods deal with. Depending on their position in the machine learning pipeline, we distinguish between:

1. Pre-processing methods
2. In-processing methods
3. Post-processing methods

Pre-processing methods follow 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. The idea is to assign different weights to the observations based on relative frequencies, so that the algorithm learns on a balanced dataset (Caton and Haas 2024).
In-Processing methods modify the optimization 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. Compared to the other two categories, the design of in-processing methods directly depends on the learner, and they require access to its internal workings. This makes them challenging to apply to black-box algorithms.
Post-processing methods are again independent of the specific learner. 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 a thresholding technique by Hardt et al. 2016, where group specific cut-offs are learned via linear optimization to re-classify the data after training.
Depending on the task (regression, classification) and the model, there are highly specified and advanced methods. For the case study in chapter 3, we limit ourselves to methods implemented in the `mlr3fairness` package.

3.4 Bias and the feedback loop

Before applying the theory to real-world data, it remains to introduce different types of biases and the context in which a machine learning model is usually embedded. Deployed

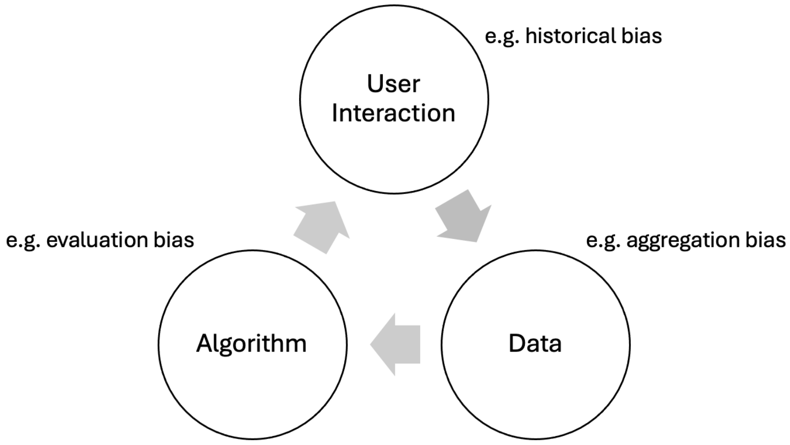

Figure 1: The *data, algorithm, and user interaction feedback loop* as described by Mehrabi et al. 2022. Different categories of bias can be introduced at each stage of the process.

as an ADM, the model assists in decisions such as whether someone gets admitted to college, receives a loan or is released from prison. It thereby indirectly contributes to shaping our reality.

Mehrabi et al. 2022 conceptualize the situation in form of the *data, algorithm, and user interaction feedback loop* (Figure 1), which can be understood as follows: As a society, we make decisions that shape our reality. The reality becomes measurable by collecting data. The algorithm learns from this data to make optimal predictions, which, in turn, inform the decision-maker's judgment. The new decisions will once again shape reality, reflected in updated data.

At each stage, bias can be introduced into the process. More dangerous, bias can even be amplified as the algorithm influences decision-making on a large scale. Consequently, every fairness project comes with the responsibility to understand the data-generating process and gain clarity on how the algorithm will be deployed in the real world.

Additionally, the *data, algorithm, and user interaction feedback loop* helps clarify which type of bias might be relevant in a given situation by placing it at a specific position in the feedback loop. Distinguishing between bias mechanisms can be crucial and should influence the definition of fairness and the choice of fairness adjustments in a given situation. This will also become evident in the following section where we examine the SQF dataset.

4 Case Study: Stop, Question, and Frisk

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. As mentioned in the introduction the so-called "New York strategy" (Gelman, Fagan, and Kiss 2007) has been criticized for disproportionally targetting African American and Hispanic individuals. This makes the recordings of the stops an interesting resource for fairness research. It also has been recommended by Fabris et al. 2022, not least in the effort to bring more diversity to the datasets used in the field. For our analysis we are interested in whether a classifier trained to predict the arrest after a stop is discriminatory with respect to race.

4.1 Setup of the fairness experiment

We compare the following models in terms of fairness and model performance, measured by the difference in true positive rates (equal opportunity) and the classification accuracy respectively:

- Regular Random Forest
- Reweighing to balance disparate impact metric (Pre-Processing)
- Classification Fair Logistic Regression With Covariance Constraints Learner (In-Processing)
- Equalized Odds Debiasing (Post-Processing)

More details about the methods can be found in Pfisterer 2024. Specifically, for Reweighing, see mlr3fairness Reweighing. Refer to Fair Logistic Regression for more details on the chosen in-processing method. For the post-processing strategy, check Equalized Odds.

4.2 Data description

As they were the most recent at the time of writing this thesis, we work with the stops from 2023. The raw 2023 dataset consists of 16,971 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 12,039 observations.[3]
We summarize "Black Hispanic" and "Black" into the group "Black" and "American Indian/ Native American" and "Middle Eastern/ Southwest Asian" into the "Other"

[3] 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 will introduce complications when we implement fairness methods.

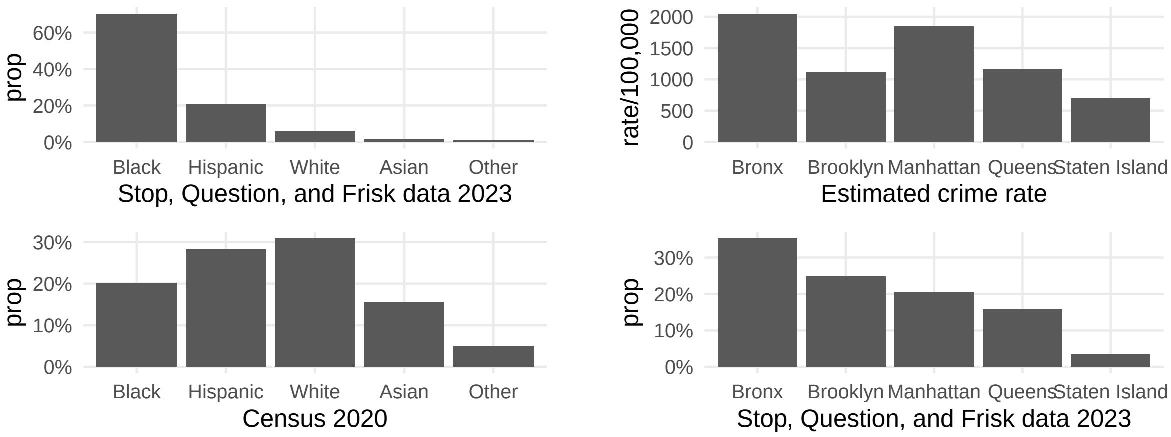

Figure 2: Bar plot comparing the distribution of ethnic groups across boroughs in the SQF 2023 and NYC from 2020 Census (left). On the right a comparison of the estimated borough-wise crime rate per 100,000 citizens with the ethnic distribution of SQF stops.

category. Black people are by far most often stopped, making up 70% of the total stops; yet, according to 2020 census data black people contribute to only 20% of the city's population (Figure 2, left). At the same time white people form the majority of New York citizens (30%) but are involved in merely 6% of the stops.

After 2012 there has been a stark decline in stops and the police is known to focus their attention on high crime areas. Therefore, we further look at each borough. The most stops in 2023 occurred in Bronx and Brooklyn. Based on report of the NYPD and population statistics from 2020, the Bronx also has the highest estimated crime rate per 100,000 citizens. Manhattan has a similarly high crime rate, but fewer stops. Note that Bronx and Brooklyn happen to be the boroughs with the highest proportion of black citizens (Figure 2, right).

After a more general overview of the dataset, we turn to the outcome of interest. In the cleaned 2023 data about 31% of stops result in an arrest. Table 3 shows that the disparities in arrestment across ethnic groups is in general low. As group fairness metrics are observational and constructed from the joint probability of $Y, \hat{Y}, A$, this already gives us a hint that the classifier trained to predict the arrestment of a suspect might show little racial disparities.

Given the evolution of stops over the years and the 2013 ruling, the question arises whether a classifier trained on data from the unconstitutional period (2004-2012) will perform differently. For a comparison, we therefore train an additional random forest classifier on data from 2011. This is the year with the most stops. We carry out the same data cleaning steps, starting with 685,724 recorded stops and reducing this to 651,567 clean observations. Note that these are around 40 times more stops than in 2023. This means that the 2011 data has substantially more low-risk stops; only around 6% result in an arrest. This is a stark contrast to the 31% in 2023. As seen in Table 4 the differences in arrestment rate across ethnic groups are minor. As in 2023 white people have the highest arrestment rate.

We select features that reflect the information available to the officer at the time of the

Group	Prop
Black	30.56%
Hispanic	31.60%
White	37.95%
Asian	37.84%
Other	31.45%

Table 3: Group-wise arrestment rates in 2023

Group	Prop
Black	5.988%
Hispanic	5.830%
White	6.859%
Asian	5.840%
Other	4.575%

Table 4: Group-wise arrestment rates in 2011

arrest decision. This includes details about the stop's progression, such as whether the person was frisked or issued a summon. We consider these outcomes as intermediate steps before an arrest. Additionally, we control for factors like the time of the stop and whether the officer was in uniform. This feature selection is inspired by Badr and Sharma 2022.

4.3 Results of the fairness experiment

For the training of the classifiers, we dichotomize the race attribute by grouping "Black" and "Hispanic" as people of colour ("PoC") and "White", "Asian", and "Other" as white ("White"). We run a five-fold cross validation and show the results in Figure 3.

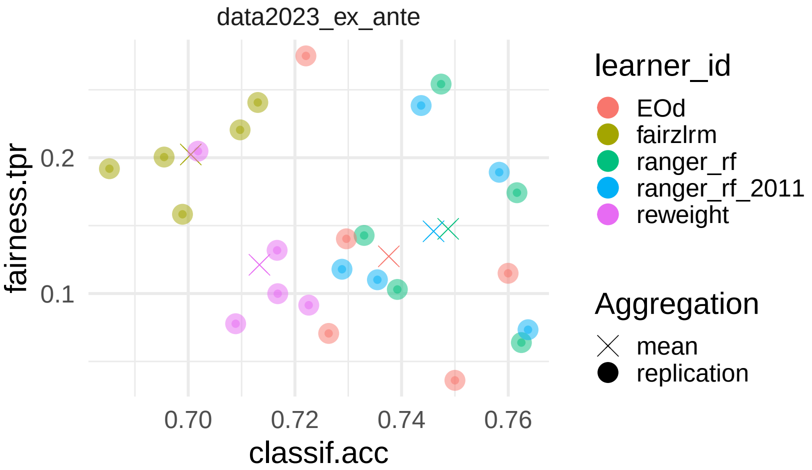

Figure 3: Comparison of learners with respect to classification accuracy (x-axis) and equal opportunity (y-axis) across (dots) and aggregated over (crosses) five folds.

The x-axis shows the learner's accuracy and on the y-axis we plot the absolute difference in true positive rates across groups. In the bottom right corner we find fair and accurate classifiers. In terms of fairness reweighing and the equalized odds post-processing method perform best. However, the regular random forest classifier comes close to their fairness performance and performs slightly more accurate. Surprisingly, it does not make any difference for the fairness if the classifier is trained on 2011 or 2023 data. We examined the model closer and find that due to the low prevalence in the population, a classifier trained on 2011 data primarily suffers from the highly skewed distribution of arrests.

The classifier largely predicts the negative label for *anyone* regardless of race, which overshadows potential fairness concerns. The fairness adjusted logistic regression performs worst in terms of accuracy and fairness. As the picture could change depending on the chosen fairness metric (y-axis), we also experimented with other metrics, such as equalized odds or predictive parity. In all cases the regular random forest does not perform worse in terms of fairness but better in terms of accuracy than most fairness adjusted classifiers. Since the classifiers perform similarly, we choose the regular random forest trained on 2023 to examine the model closer. In Figure 4 on the left we plot the prediction score densities for each group. We can see that in general white people tend to have higher predicted probabilities than PoC. The mode for the scores for non-white individuals is around 0.05 while it is around 0.125 for white individuals. The score resembles the probability of being predicted positive (arrested). This may suggest that PoC are involved in more low-risk stops than white suspects. In Figure 4 on the right we plot the absolute difference in selected group fairness metrics. Exact equality of the group fairness metrics cannot be expected in practice, so it is common to allow for a margin of error ϵ. Taking $\epsilon = 0.05$, the classifier is fair according to each of the selected metrics, though the difference in positive predictive rates is close to 0.05.

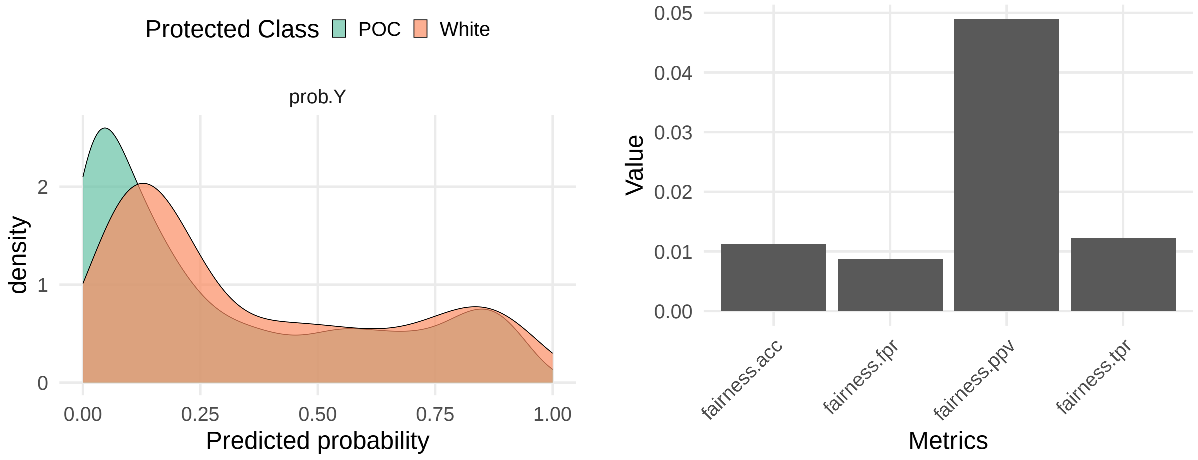

Figure 4: Fairness prediction density plot (left) showing the density of predictions for the positive class split by ”PoC” and ”White” individuals. The metrics comparison barplot (right) displays the model’s absolute differences across the specified metrics.

For a more nuanced picture, we additionally report the group-wise error metrics in Table 5. As the absolute differences in the plots have already shown, the *separation* metrics are virtually the same across groups. For the *sufficiency* metrics, the positive predictive value is slightly higher for white individuals, meaning a positive prediction is more likely to be correct for them than for PoC. Conversely, the negative predictive value is higher for PoC, indicating that a negative prediction is more likely to be correct for them than for white individuals.

In general, it seems like the classifier trained on SQF data to predict the arrest of a suspect is not discriminatory against PoC. In contrast, it even performs slightly better for PoC than for white people on many of the common performance metrics. Badr and Sharma 2022 have similar findings.

	TPR	FPR	NPV	PPV	FDR	Acc
PoC	0.75	0.07	0.89	0.84	0.16	0.88
White	0.74	0.06	0.85	0.89	0.11	0.86

Table 5: Group fairness metrics for RF classifier trained on 2023 SQF data.

In their study they chose six representative machine learning algorithms (Logistic Regression, Random Forest, Extreme Gradient Boost, Gaussian Naïve Bayes, Support Vector Classifier) to predict the arrest of a suspect. Fairness is measured with six different metrics (Balanced Accuracy, Statistical Parity, Equal Opportunity, Disparate Impact, Avg. Odds Difference, Theil Index) and separate analysis are conducted with sex and race as PA. They compare the fairness of the regular learner to the fairness of learner with a pre-processing method (reweighing) and a post-processing method (Reject Option-based Classifier). All in all, they find that the regular models to not perform worse in terms of fairness than the fairness adjusted models. This leads them to conclude ”[...] that there is no-to-less racial bias that is present in the NYPD Stop-and-Frisk dataset concerning coloured and Hispanic individuals.”

What both of our case studies have in common is that the models were trained on recent data and select arrest as target. We trained our model on 2023 stops and Badr and Sharma 2022 used 2019 stops. The authors specifically identified time as an important factor in their results and state: ”The NYPD has taken crucial steps over the past years and significantly reduced racial- and [gender-based] bias in the stops leading to arrests. This conclusion nullifies the common belief that the NYPD Stop-and-Frisk program is biased toward [coloured] and Hispanic individuals.” Is this the whole picture?

5 Studies on the SQF Dataset

5.1 Approaches to fairness in SQF

Before going into detail about a specific study, we provide a tabular overview of the different approaches to fairness in the SQF data. We will go into more depth into two of them in the following.

Authors	Task	Model	Fairness Metric	Results
Kallus and Zhou 2018	Predict prob. of innocence (no weapon)	Log. Regression	Equal Opportunity, Equalized Odds	Bias against PoC
Rambachan and Roth 2016	Possession of contraband	Log. Regression	No explicit fairness metric; evaluate prediction function properties	No bias against PoC
Badr and Sharma 2022	Predict probability of arrest	Log. Regression, RF, XGBoost, GNB, SVC	Balanced Accuracy, Stat. Parity, Equal Opportunity, Disparate Impact, Theil Index	No bias against PoC
Khademi et al. 2019	Predict probability of arrest	Weighted regression models	FACE causal fairness (group), FACT fairness (individual)	No group bias, but individual bias
Goel, Rao, and Shroff 2016	Predict possession of weapon	(Penalized) Log. Regression	No explicit fairness metric; group-wise hit rates	Bias against Black and Hispanic

Table 6: Overview of SQF-related fairness studies. This table summarizes findings from five key studies evaluating the fairness of the stop-and-frisk policy. Depending on the task and model used, the studies reach different conclusions.

One of the main challenges with the NYPD's data is that, when evaluating the fairness of stop-and-frisk as a policing strategy, various tasks can be formulated to address the question. Only some of them are suitable to make conclusions about the fairness of the stop-and-frisk policy as a whole.

Like Badr and Sharma 2022, we trained a classifier to predict arrests and used group metrics to assess fairness. However, given that we also trained a ”fair” classifier on data from 2011, but the stop-and-frisk practice was officially declared unconstitutional for 2004 to 2012, the task we (and Badr and Sharma 2022) chose, is not a reliable indicator of the overall fairness of the policy.

To answer the question of fairness in stop-and-frisk other studies take a step back and identify a problem with how the data is generated. They formalize and acknowledge that the discrimination in SQF does not solely lie in the outcome of the stop but the decision to stop someone in the first place.

5.2 Residual unfairness

In their paper **Residual Unfairness in Fair Machine Learning from Prejudice Data** Kallus and Zhou 2018 conceptualize the problem as shown in Figure 5. A person is defined 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 ($Z = 1$) or not ($Z = 0$). This mechanism can be seen as a category of selection bias and, referring back to the feedback loop (Figure 1), is introduced by the user.

In the SQF context, we can imagine that the police are generally more suspicious towards PoC than white people, resulting in more stops of the former. Alternatively, one could argue that they are more likely to stop individuals in high-crime areas, which happen to be mostly low-income neighbourhoods largely populated by PoC.

Naturally, we can only know the outcome $Y \in \{0, 1\}$ of a stop for the individuals who were stopped. This can create a situation in which the training data produced by the biased decision policy is not be representative for the population the algorithm will be deployed on. 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.

The problem for fairness in this case is that fairness adjustments of the learner (trained on the biased sample) do not translate to its deployment on the target population. Even fairness-adjusted classifiers can discriminate against the same group that has historically faced discrimination (Kallus and Zhou 2018). They call the remaining disparities in fairness metrics after fairness adjustments have been made **residual unfairness**.

At this point, we refer back to Figure 2. The left plot shows a clear difference between the racial distribution in the SQF data and the city as a whole. In terms of race, the sample is clearly not representative for NYC[4]. At the same time the estimated borough-specific crime rates also differ from the distribution of stops per borough as seen in the right plot. Kallus and Zhou 2018 demonstrate that their theoretical findings are reflected in the SQF data. Their task is to predict a person's innocence. They define innocence as not carrying an illegal weapon and guilt as carrying one. The reasoning behind this approach is that the discriminated group is the one more frequently falsely accused of possessing an illegal weapon. The authors find that non-white individuals are indeed more often wrongfully classified as guilty. Even after applying a post-processing strategy to achieve equalized odds, unfairness against PoC persists when the classifier is used on NYC's overall target population.

[4] It can be questioned whether it makes sense to require the SQF sample to be representative for the population of NYC. It might make more sense to require that it is representative of the population of *criminals* in NYC.

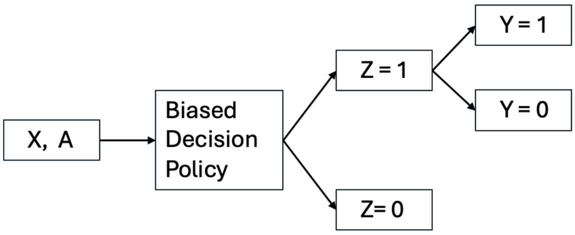

Figure 5: Selection bias in the SQF data, as conceptualized by Kallus and Zhou 2018. The true label is only known for the stopped individuals ($Z = 1$).

5.3 Bias in, bias out?

Another perspective is offered by Rambachan and Roth 2016 in their paper **Bias In, Bias Out? Evaluating the Folk Wisdom**. While the main message of Kallus and Zhou 2018 is that even fairness adjusted classifiers exhibit the "bias in, bias out" mechanism Rambachan and Roth 2016 argue that it depends on the chosen classification task.
Similar to Kallus and Zhou 2018 they are interested in whether a person carries a contraband $Y \in \{0, 1\}$. 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. More precisely, they see the biased-decision policy in the decision to *search* someone.
When we previously defined the biased selection mechanism as $Z = 1$ corresponding to stop and $Z = 0$ corresponding to no stop, this study sees the biased decision policy in the decision to search someone, which then should be denoted as $Z^* = 1$ for search and $Z^* = 0$ for no search. Though the studies select different variables for the biased decision policy their conceptual frameworks remain comparable. Again, only on searched people a contraband can be found. The goal is to estimate the possession of a contraband Y, but we estimate this from $Y|Z^* = 1$.
In contrast to Kallus and Zhou 2018, the authors of this paper argue that the classifier shows the opposite effect; instead of continuing to discriminate the previously disadvantaged group, the classifier exhibits *less* bias as the prejudice against African-Americans increases.
As the police becomes more biased towards African-Americans, they search them more leniently. This means that many innocent African-Americans are included in the searched observations. Consequently, the model learns on average lower risk scores for African-Americans. Essentially, the data for African-Americans becomes more noisy, which lowers the predicted probabilities for this group. The authors call this mechanism **bias reversal**.
Rambachan and Roth 2016 continue to examine alternative classification tasks that could be constructed from the SQF data. In their second scenario, they train an algorithm to predict whether to search someone in the first place ($Z^* \in \{0, 1\}$). Now the search becomes the target and in this case **bias inheritance** is observed, meaning the classifier continues to discriminate the historically disadvantaged group.
The same happens for a two stage classification task that first predicts whether to search someone and if they searched someone if the individual carries a contraband ($Y \cdot Z^* \in$

$\{0,1\}$). What happens here is that as more of the stopped African Americans are also searched, the algorithm learns to associate search with more with African Americans than with white people and thus in the future also predicts higher probabilities for a search for African Americans. This is the **bias inheritance** mechanism. We can see parallels to this paper to our own case study in the sense that, PoC indeed have lower risk scores (Figure 4, left) and are relatively speaking less often predicted as arrested as white individuals.
As seen in Table 6 there are more studies that have worked with the SQF data, each with a unique approach to the question of fairness. Khademi et al. 2019 are also interested in whether the decision to arrest an individual, after a stop has been made, is discriminatory with respect to race. They design two causal fairness methods, namely the Fair on Average Causal Effect (FACE) and the Fair on Average Causal Effect on the Treated (FACT), to estimate the causal impact of race on the outcome. Their notions of fairness are based on counterfactual reasoning, utilizing Inverse Probability Weighting and Matching to estimate the quantities of interest. While one of their metrics finds that the odds of being arrested after a stop are higher for Black-Hispanics than for white individuals, the other metric does not show any racial discrimination.
Goel, Rao, and Shroff 2016, on the other hand, focus on the prediction of the possession of a weapon. They construct a statistical model to account for location specific crime rates and developments of criminal activity over time. After controlling for these factors, the authors conclude that Black and Hispanic individuals are disproportionately often involved in low-risk stops.

6 Conclusion

The SQF data presents interesting questions and challenges. In our examination of fairness in SQF, we placed particular emphasis on selection bias and its impact on the fairness assessment. Future studies could explore other forms of bias that may be relevant to this dataset and examine how they influence the fairness of classifiers. One notable example is historical bias, which may play a significant role here.

Moreover, it would be valuable to investigate if and how the fairness audit changes, if race is not aggregated into a binary PA but is instead treated a multi-class variable. This approach could lead to different fairness conclusions and provide a more nuanced perspective. However, in this context, class imbalance within the protected attribute would become more pronounced, introducing new research challenges.

Through our work, we have demonstrated that before implementing any fairness intervention, it is essential to first formulate a clear and well-defined fairness question. There is a fundamental difference between asking whether stop-and-frisk as a whole is fair or whether a classifier trained to predict the arrest of a person is fair. The way we frame the question can shape the design of entirely different algorithmic tasks and fairness analyses. After conducting a detailed fairness audit, experimenting with various models, and reviewing multiple studies, our answer to ”Is the Stop, Question, and Frisk practice fair?” remains: it is complex.

List of Figures

List of Tables

A Electronic Appendix

See the GitHub repository for data, code and illustrations: SQF Fairness Project

References

Badr, Youakim and Rahul Sharma (June 2022). “Data Transparency and Fairness Analysis of the NYPD Stop-and-Frisk Program”. In: *Journal of Data and Information Quality* 14.2, pp. 1–14. ISSN: 1936-1955, 1936-1963. DOI: 10.1145/3460533. (Visited on 12/24/2024).

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Binns, Reuben (2020). “On the apparent conflict between individual and group fairness”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, pp. 514–524. ISBN: 9781450369367. DOI: 10.1145/3351095.3372864. URL: https://doi.org/10.1145/3351095.3372864.

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).

Chouldechova, Alexandra (2016). “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. In: *Big data* 5 2, pp. 153–163. URL: https://api.semanticscholar.org/CorpusID:1443041.

Corbett-Davies, Sam et al. (Jan. 2023). “The measure and mismeasure of fairness”. In: *J. Mach. Learn. Res.* 24.1. ISSN: 1532-4435.

Fabris, Alessandro et al. (Sept. 2022). “Algorithmic fairness datasets: the story so far”. In: *Data Mining and Knowledge Discovery* 36.6, pp. 2074–2152. DOI: 10.1007/s10618-022-00854-z. URL: https://doi.org/10.1007%2Fs10618-022-00854-z.

Favier, Marco et al. (Dec. 2023). “How to Be Fair? A Study of Label and Selection Bias”. In: *Machine Learning* 112.12, pp. 5081–5104. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-023-06401-1. (Visited on 02/05/2025).

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).

Gelman, Andrew, Jeffrey Fagan, and Alex Kiss (Sept. 2007). “An Analysis of the New York City Police Department’s “Stop-and-Frisk” Policy in the Context of Claims of Racial Bias”. In: *Journal of the American Statistical Association* 102.479, pp. 813–823. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214506000001040. (Visited on 01/08/2025).

Ghani, Rayid et al. (2025). *Chapter 11: Bias and Fairness — Big Data and Social Science*. (Visited on 01/15/2025).

Goel, Sharad, Justin M. Rao, and Ravi Shroff (Mar. 2016). “Precinct or Prejudice? Understanding Racial Disparities in New York City’s Stop-and-Frisk Policy”. In: *The Annals of Applied Statistics* 10.1. ISSN: 1932-6157. DOI: 10.1214/15-AOAS897. (Visited on 11/19/2024).

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).

Khademi, Aria et al. (May 2019). “Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality”. In: *The World Wide Web Conference*. San Francisco CA USA: ACM, pp. 2907–2914. ISBN: 978-1-4503-6674-8. DOI: 10.1145/3308558.3313559. (Visited on 12/24/2024).

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2017). “Inherent Trade-Offs in the Fair Determination of Risk Scores”. In: *LIPIcs, Volume 67, ITCS 2017* 67, 43:1–43:23. ISSN: 1868-8969. DOI: 10.4230/LIPICS.ITCS.2017.43. (Visited on 02/27/2025).

Lakkaraju, Himabindu et al. (Aug. 2017). “The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: ACM, pp. 275–284. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098066. (Visited on 12/25/2024).

Makhlouf, Karima, Sami Zhioua, and Catuscia Palamidessi (May 2021). “On the Applicability of Machine Learning Fairness Notions”. In: *ACM SIGKDD Explorations Newsletter* 23.1, pp. 14–23. ISSN: 1931-0145, 1931-0153. DOI: 10.1145/3468507.3468511. (Visited on 12/01/2024).

Mehrabi, Ninareh et al. (July 2022). “A Survey on Bias and Fairness in Machine Learning”. In: *ACM Computing Surveys* 54.6, pp. 1–35. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3457607. (Visited on 01/07/2025).

Pfisterer, Florian (2024). “Algorithmic Fairness”. In: *Applied Machine Learning Using mlr3 in R*. Ed. by Bernd Bischl et al. CRC Press. URL: https://mlr3book.mlr-org.com/algorithmic_fairness.html.

Rambachan, Ashesh and Jonathan Roth (2016). *Bias In, Bias Out? Evaluating the Folk Wisdom*. URL: https://drops.dagstuhl.de/storage/00lipics/lipics-vol156-forc2020/LIPIcs.FORC.2020.6/LIPIcs.FORC.2020.6.pdf.

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).

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.
AI was used in the following ways in this work: GitHub Copilot, was used to help writing code, and is integrated into my RStudio and Visual Studio Code environment. Literature research with ”Search” tool in ChatGPT-3.5, Elicit and Connected papers. Occasional check in understanding of concepts with ”Reason” tool in ChatGPT-3.5. Did not take this to generate new ideas but to validate my own understanding. Improvement of language and grammar with ChatGPT-3.5. Explicitly stated to keep the meaning and content of the text. Assistance of ChatGPT-3.5 and GitHub for some latex code (e.g. the tables, insertion of images).

Munich, February, 28th 2025

Juliet Fleischer