On Algorithmic Fairness and Bias Mitigation in Recidivism Prediction

An econometric view on observed tradeoffs between conflicting definitions of fairness, and the applicability of post-processing methods for bias correction of criminal sentencing algorithms

by Lennert Jansen[1]

Under the supervision of
dr. Paul Bouman[2] & Benjamin Timmermans.[3]

Winter, 2020

Abstract

Ensuring fair treatment of historically disadvantaged groups of individuals by machine learning (ML) guided decision-making systems is a rapidly growing point of discussion in both academics and commercial industries. This thesis aims to investigate whether a popular recidivism prediction instrument (RPI), known as COMPAS, can be accused of being unfairly biased against African-Americans and/or women. Furthermore, the applicability of certain bias mitigation post-processing algorithms is studied for debiasing an arbitrary probabilistic recidivism predictor. Statistically conclusive results suggest that COMPAS-scores are in fact unfairly putting African-Americans at a disadvantage. However, the results with respect to a bias against women are inconclusive. Finally, reject option based classification (RObC) proves highly effective for achieving group-based fairness optima, while preserving balanced accuracy. However, these group-based fairness measures are optimised at the expense of an arguably important fairness notion, known as calibration.

[1]499715lj@eur.nl
[2]bouman@ese.eur.nl
[3]b.timmermans@nl.ibm.com

Contents

Acknowledgements

The process of writing my thesis has been anything but dull: from having to abandon my original use-case three months into my internship, to interesting video-calls with IBM researchers in the United States, to balancing the completion of this paper with a new master's degree and a part-time job. It never lacked in excitement, albeit from time to time of the stress-inducing variety.

Coming into the project, I had no background whatsoever on the subject of fairness in machine learning. I also had to Google the word 'recidivism'. But after sifting through a healthy amount of papers, watching hours of videos on YouTube, and long conversations Zoltan, Tim, and others at IBM, I started to get an idea of what I was doing. It then developed towards a deep interest in the subject. And I would be delighted if I can continue to work on the subject in the future.

I would like to thank my supervisors Paul Bouman and Timmermans for their guidance during the process of writing this thesis, and giving me the freedom to complete my work alongside my other responsiblities. I would especially like to express my gratitude to my parents, Harry and Cyrillia, and my two brothers, Max and Karel. During the long and challenging process of my graduation research, they were undoubtedly my most frequently called upon, and reliable support-system. My colleague and desk-neighbour at IBM, Tim, was a wonderful sparring partner when we started our respective graduation projects. The numerous walks we went on after lunch always managed to spawn motivational and intriguing topics of conversation. My roommate, Wout-Jan, also deserves a mention of appreciation for having patiently dealt with my inconvenient departure- and arrival-patterns, and the piles of scientific articles and printed drafts that occasionally cluttered the kitchen table. Finally, I would also like to explicitly thank my cousin Jolanda, and my friends Emile, Michael, Philip, Comethazine, Morena, Joris, Floris, Floris K., Bram, Coen, Dirk, Isadora, Maxine, and Joshua.

March 11, 2020
Amsterdam, the Netherlands

1 Introduction

Many U.S. courtrooms use state-of-the-art statistical software, like that of Northpointe Inc. (now Equivant) to estimate the likelihood of a defendant becoming a recidivist, a term used to describe convicted criminals who reoffend. Alarmingly though, a recent study shows that Northpointe's decision support tool is almost twice as likely to mislabel black defendants as recidivists. Northpointe's COMPAS (an acronym for Correctional Offender Management Profiling for Alternative Sanctions) is arguably the most widely used tool of its kind, with standard use in various phases of the criminal justice process in Broward County of Florida, the states of Wisconsin, New York and California, to name a few. COMPAS produces a risk score on a decile scale, based on a variety of variables, including a defendant's criminal history, degree of offence, family's criminal history, age, gender and opinions on various societal issues believed to be related to relevant criminological indicators. These risk scores are then taken into account when, for instance, determining a defendant's bail payment, risk of pretrial misconduct or likelihood of reoffending upon release.

Subsequently, the introduction of a sentencing reform bill in the U.S. Supreme Court, that mandates the use of such risk assessment tools in all the nation's courtrooms, inclined a New York-based nonprofit investigative newsroom named ProPublica to place the COMPAS recidivism prediction instrument (henceforth RPI) under harsh scrutiny. Their 2015 study into the suspected unfair treatment of African-American defendants by COMPAS spawned a heated debate and numerous follow-up studies about Northpointe's statistical tool and algorithmic fairness in general (Dieterich, Mendoza, & Brennan, 2016; Feller, Pierson, Corbett-Davies, & Goel, 2016).

In general, an increasing number of high-stakes decisions are being made about individuals by artificially intelligent (AI) systems, significantly impacting their lives and communities alike. As people continue to rely on algorithmically guided decision-making, societies must prioritise ensuring that these models align with their norms and values. That is why recent years have witnessed a noticeable increase in the emphasis on fairness, accountability and transparency in machine learning literature. Not only academia, but prominent tech companies too are investing growing portions of their time and funding in search of methods for identifying and correcting unethical systematic disparities induced by decision-making models, known as algorithmic biases.

International Business Machines Corporation (hereafter, IBM) is no exception, having recently deployed their AI Fairness 360 (hereafter AIF360) toolkit, an open-source Python library, dedicated to helping users detect, understand and mitigate unwanted algorithmic bias (Bellamy et al., 2018). This extensible toolkit serves as a platform on which commercial industry data scientists and machine learning academics can exchange and evaluate bias correcting algorithms. Furthermore, it is also the main resource of fairness related methodology for this study.

1.1 Research goals and hypotheses

The aim of this thesis is to investigate to what extent Northpointe's criminal risk assessment tool, COMPAS, can be accused of being unfair with respect to race and gender. Additionally, this research aims to study the effectiveness of various algorithmic bias mitigation techniques as methods for correcting potential unfairness. Based on previous work on the subject, it is hypothesised that unfair biases in disfavour of African-Americans and males. However, these disparities are expected to be more moderate than, for instance, claimed by Angwin, Kirchner, Mattu, and Larson (2016), due to methodological limitations of their study, outlined in Section 3. Furthermore, certain discrepancies in recidivism prediction efficacy for males is expected to be attributable to demonstrable differences in (re)offence prevalence between men and women. The presumed racial and potential gender biases

are quantified and tested using both conventional econometric methods, as well as contemporary individual- and group-based fairness metrics.

1.2 Outline of thesis

The remainder of this paper is structured as follows: the next section contains an extensive theoretical background on discrimination, algorithmic fairness in decision support tools for criminal justice, algorithmic bias and fairness, and important trade-offs between non-discrimination and prediction accuracy. Section 3 provides a brief overview of related work, previously done on the subject of algorithmic fairness in criminal justice decision-support modelling, including a brief summary of ProPublica's accusation of COMPAS' discriminatory properties and the most important follow-up studies. Then, Section 4, covers IBM's open-source fairness Python toolkit, along with the methods considered for this research. The most important results obtained are presented in Section 5. Finally, Sections 6 and 7 contain a recapitulation of this study's points of discussion and most important conclusions, respectively.

2 Background

2.1 Fairness in Machine Learning

Mehrabi, Morstatter, Saxena, Lerman, and Galstyan (2019) define fairness in the context of decision-making and machine learning as follows: "*In the context of decision-making, fairness is the absence of any prejudice or favouritism toward an individual or a group based on their inherent or acquired characteristics.*" These characteristics are described in more detail in Section 2.5.1.

But why should we care about embedding societal norms and values into decision-making algorithms? In short, ensuring fairness is upheld by socio-technical systems will likely unfold to the collective benefit of the societies in which they operate.

Since the industrial revolution, machines started fulfilling human tasks. What began as a means to perform repetitive and exhausting *manual* labour, is now transitioning *mental* tasks that are either too repetitive, prone to human error, or simply impossible for human brains. Automation has moved from our hands to our brains. Consequently, as more and more decisions are being made by algorithms, and these programmes become more sophisticated, the amount of influence these choices exert on our lives increases. As alluded to in Section 1, machines in this day and age make choices about hiring, prosecution, police patrolling, college admissions and mortgage applications, among other matters. These decisions have significant impact on the lives of individuals and their communities.

There is heightened concern in the scientific community and growing attention in public media about the demographic disparities that might arise from this phenomenon (Narayanan, 2018). Originally, the aforementioned decisions used to be made by humans, often guided by their biases, histories, prejudice and instinct. Especially during less emancipated and more racially skewed times in the not-so-distant past, minority groups were systematically disadvantaged. Nowadays, the same choices are made by computer programmes, trained on data recorded and gathered by humans. Although these algorithms are designed as facially neutral and objective systems, they may still reflect our human biases, thus perpetuating or exacerbating societal disparities (Barocas, Hardt, & Narayanan, 2018). The data on which these models are learned are a reflection of humanity. And albeit by malicious design of stakeholders, or contaminated training data originating from an ethically unjust past, algorithmic discrimination is an observed problematic outcome and should be dealt with promptly.

Fairness has become such a hot-button issue that a yearly conference has been fully dedicated to the research area (ACM Conference on Fairness, Accountability and Transparency). This inter-disciplinary nascent field of study that draws from computer science, statistics, law, psychology and economics, is primarily concerned with ensuring non-discrimination in socio-technical systems. And although it is in full bloom, there is still no consensus on appropriate means of alleviating algorithmic unfairness (Hardt, Price, Srebro, et al., 2016).

2.2 Discrimination: from legal doctrine to algorithmic constraints

Understanding discrimination is crucial to effectively ensuring equitable treatment by socio-technical systems. When is discrimination wrong? What does it mean to discriminate against an individual or a group of people? It may come as no surprise that these and related questions cannot be answered by a one-size-fits-all rebuttal. Barocas et al. (2018) argue that discrimination is not a general problem. Rather, it is domain and feature specific. Altman (2016) acknowledges this statement by postulating that there is no universally accepted definition of discrimination. Moreover, the same publication finds that, despite discrimination being outlawed by six of the core human rights documents, these very treaties do not define discrimination at all. They simply present a non-exhaustive list of attributes on the basis of which discrimination is prohibited. Due to the nature of algorithmic design, one cannot robustly reduce socio-technical discrimination without properly defining it. Luckily, when focusing on specific domains, formalisations of this illusive notion arise, resulting in various societal, lawful and economic definitions of discrimination. d'Alessandro, O'Neil, and LaGatta (2017) suggest that fairness should be embedded in ML systems by finding the most relevant legal or economic notion of discrimination, given the application and context, then proposing an appropriate best-fit metric.

The attentive reader might wonder if discrimination isn't the very point of machine learning. ML systems essentially try to find patterns in data on the grounds of which lines can be drawn between groups for classification and prediction. Then why are we accusing the machine learning community of malicious practice, while these algorithms are doing exactly what they have been designed to do? Clearly, this thesis concerns a different kind of discrimination. Discrimination in machine learning is perfectly permissible, unless there is an unjustified basis for it, practical or moral irrelevance. Inadvertently disadvantaging protected groups can also be sufficient reason for policy-makers and auditors to scrutinise a decision-making system.

As mentioned before, international legal doctrine provide myriad ways of defining discrimination. We therefore focus on judicial framings of discriminatory practice involved in ranking systems or classification (e.g., hiring, housing, lending, recidivism assessments, etc.). The two prevailing theories of liability that are most appropriate for these kinds of problems are named *disparate treatment* and *disparate impact* (Barocas & Selbst, 2016). d'Alessandro et al. (2017) succinctly describe the former as differential treatment on the grounds of membership in a protected group, leading to disadvantageous outcome for members of that class. Disparate treatment, for instance, covers the blatant denial of opportunities based on group membership, irrespective of whether or not considering the sensitive attribute increases utility. Note how for disparate treatment, intent is of more importance than discriminatory effect. Namely, if a malicious employer intentionally considers ethnicity in a hiring model, but the model deems the variable unimportant, consideration of race will lead to few disparities. The stakeholder, however, is still in violation of disparate treatment, despite his policy not harming a protected group.

Disparate impact alludes to practices that are facially neutral or benign, yet result in disproportionately unfavourable impact on a protected class. In other words, a policy can be seen as

explicit intent (Corbett-Davies & Goel, 2018). The term was coined in the famous U.S. Supreme Court case *Griggs v. Duke Power Company* (1971). Duke Power Company required applicants for higher paying jobs to be in possession of a high school diploma, which at the time significantly reduced the number of eligible black applicants. However, the Supreme Court ruled that the company's requirements were irrelevant for an applicant's ability to perform the job, and thus found Duke Power Company guilty of discriminating against blacks. It is worth mentioning that disparate impact law solely prohibits unjustified differential outcome. Namely, if the Supreme Court had found that possession of a high school diploma was in fact essential for eligibility, it would have ruled in favour of Duke Power Company and the disparate outcome would be legal (Corbett-Davies & Goel, 2018).

In the context of algorithmic fairness, disparate treatment is of lesser importance. Corbett-Davies and Goel (2018) point out that the primary concern is whether socio-technical systems unintentionally lead to disparities, albeit due to malicious design or implicit biases embedded in the data on which they are trained. Therefore, our focus will be primarily on disparate impact, rather than treatment.

2.3 Algorithms, Bias and Criminal justice

Model-based risk assessments are well-established parts of the modern prosecution process in at least 44 countries (Singh et al., 2014). Examples of use-cases in different countries include court-ordered hospitalisation for long-term treatment ("terbeschikkingstelling" or TBS) in the Netherlands, preventive detention in Canada and sex-offender civil commitment in the U.S. (Blais, 2015; Fabian, 2012; van Marle, 2002). Pretrial risk of misconduct, bail amounts and likelihood of general, violent or sexual recidivism are commonly determined in part by some *facially* neutral algorithm. These scores are based on input from criminological and mental health professionals, given to judges, prosecutors, police and probation officers, and ultimately affect both the defendant's freedom and public safety. It is therefore important that these risk scores are reliable predictors, as mislabelling a defendant as a risky criminal (false negative error) could result in the unjust incarceration of an innocent individual, whereas erroneously deeming a dangerous criminal low risk (false positive error), thereby releasing said delinquent into society, could infringe public safety.

The goal of the small but growing number of academics is investigating how machine learning can improve and help to understand decision-making in criminal prosecution (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2017). However, the task of comparing algorithms and judicial professionals is difficult, because the data on which models are trained are produced by the judges and probation officers themselves. The problem of counterfactual outcomes further complicates the task. Namely, there is only rearrest data about released defendants, whereas counterfactuals (i.e., the impossible observation of a detained defendant *not* reoffending) have to be estimated, thereby introducing uncertainty. The matter becomes more convoluted when realising that judges take into account broad set of (often difficult to quantify) variables such as severity of crime, racial inequalities, impact on families and communities. Whereas algorithms are solely concerned with straightforward outcome variables and are sometimes less nuanced by construction.

Criminal risk assessment tools are typically constructed using econometric models of weighted factors, suggested by criminological research to be predictive of future delinquent behaviour. Risk factors such as the defendant's number of prior offences, family's criminal history, gang affiliation, substance abuse and mental health are assigned numerical values and compiled according to domain experts' knowledge to produce a risk score. Not only *static* criminal risk factors are considered, but *dynamic* factors too are often taken into account by risk assessment instruments. Dynamic factors refer to changeable characteristics such as pro-criminal attitude, criminal personality, recreational

The use of risk scores is motivated by the inaccuracy and susceptibility to bias of purely human-made decisions. Here, bias defined as "*a systematic error in reasoning or logic that occurs as the result of the automaticity with which the human mind processes information based on expectations and experience*" (Tversky & Kahneman, 1974). In one psychiatric study, it is shown that unstructured professional judgements about recidivism risk of mentally ill defendants had a false negative rate of nearly 70% (Monahan, 1982). Furthermore, there is ample evidence that courtroom decisions made by professional expertise are often associated with a defendant's race or gender, resulting in the unfair treatment of minority groups (Everett & Wojtkiewicz, 2002). Criminal justice experts and statisticians have therefore made significant efforts to devise structured estimates about criminal risk, with the goal of reducing error rates and prejudice in prosecution. In a broader sense, the goal of using data-driven risk assessment tools is not only removing human subjectivity, but also lowering crime and incarceration rates without affecting public safety (Stevenson, 2018). Advocates of evidence-based risk assessment instruments therefore argue that the use of such tools will rid criminal sentencing of human inefficiencies, diminish prison populations and maintain societal well-being.

However, sceptics of the criminal risk assessment trend fear that these criminal sentencing technologies systematically disfavour minority groups, thereby perpetuating or sometimes even exacerbating societal disparities. Recent studies resonate these concerns by suggesting that these risk assessment tools often fail to fully remove unfair treatment on the basis of stereotypes (Angwin et al., 2016; Stevenson, 2018). Criminal risk assessment instruments have been accused of exacerbating unlawful disparities among traditionally disadvantaged communities, such as ethnic minorities and women. These allegations were first voiced in 2014 by the former U.S. Attorney General, Eric Holder, and formally investigated for the first time by ProPublica in 2015. Their widely read research found, among other violations of fairness, COMPAS scores, a popular risk assessment tool in U.S. courtrooms, to be about twice as likely to mislabel African-American defendants as high risk (false positive) than it would Caucasians. ProPublica's landmark article also sparked a series of follow-up and refuting studies, the timeline and details of which are presented in Section 3.

2.4 On the origin of biases

As a technical matter, bias is something most scientists are familiar with. In statistics, the word could refer to the bias of an estimator. A psychologist probably recalls cognitive human biases, like confirmation, hindsight, survivorship or selection bias. Machine learning engineers are typically concerned with an algorithm's set of rules it uses to classify previously unseen observations, known as inductive bias. Despite various nuances, 'bias' generally refers to a systematic error or discrepancy. Of course, these technical notions of bias can also raise societal implications (Barocas et al., 2018). However, the bias that provokes the concerns mentioned in Section 2.1 is slightly different. Above all, it is an ethical issue. We shall call this notion, unwanted algorithmic bias. Bellamy et al. (2018) succinctly describe bias in the context of fairness as an unwanted systematic error that places privileged groups at a systematic advantage and unprivileged groups at a systematic disadvantage.

But what causes this bias? Despite the scientific community's abundance of definitions of socio-technical discrimination, there is still much work to be done on understanding the intricate processes that give rise to algorithmic bias. This subsection discusses potential mechanisms that lead to unwanted model bias. These postulated causes of bias are elaborately listed and discussed by Barocas and Selbst (2016). Hence, this subsection cites the their publication as its main source.

2.4.1 Target variables versus class labels.

The outcomes of interest that machine learning models try to predict are known as target variables, and thus defines what the data miner is trying to find. Whereas class labels are the mutually exclusive categories in which all possible values of the target variable can be divided. Note that a target variable is a machine interpretable abstraction or representation of an entity of interest for a model-builder. Defining a proper target variable is a non-trivial and highly subjective task. And it is through this open-ended procedure of attempting to properly construct a target variable that data miners may inadvertently define it, such that it opens the door to systematic impairment of individuals (Barocas & Hardt, 2017; Barocas & Selbst, 2016).

Take a credit scoring example. Creditworthiness is not a measurable real-world entity. Rather, it is the subjective perception of a person's likeliness fulfil the duties required by the credit industry's own system. Hence, creditworthiness is a device of the problem itself, yet it is taken as the standard by which companies decide whether or not to extend loans. Furthermore, there is evidence to believe that the current compositions of credit scores disadvantage minority groups (Hurley & Adebayo, 2016; Waddell, 2016). The problem of exerting bias through inappropriate defining of target variables is not limited to credit scoring. Consider, for instance, how one should quantify what makes a "good employee". Is it expected tenure or sales? Barocas and Selbst (2016) also show how certain definitions of target variables for desirable employee traits can cause disparate impact. Finally, the same holds for recidivism scores (Chouldechova, 2017). Angwin et al. (2016), for instance, have shown how current scoring systems used in U.S. criminal sentencing have intrinsic tendencies to disproportionately disadvantage blacks.

2.4.2 Training data.

Machine learning models learn by example. And a model is only as fair as the data it has been trained on. Hence, if the model's training data is tainted with or by human bias, naturally, the model's output will reflect it. The "with or by" part of the previous sentence hints at the two ways training data can be the cause of unwanted algorithmic bias. Either when human prejudice in past decisions have led to biased training examples entering the data set, and these biased examples are seen as valid observations to learn from, or human bias in sample selection has led to over-representation of one group, leading to systematic disadvantages for the under-represented subpopulation (Barocas & Selbst, 2016).

The first case, known as *skewed samples*, arises from the process of manually labelling examples by assigning them class labels (Barocas & Hardt, 2017). Think of constructing training data for a hiring algorithm using examples of past résumés that were invited for interviews. If this recruitment process was unjustly guided by a preference for male applicants in the past, training a model on these examples will perpetuate that bias (Dastin, 2018).

In a more complex example, Lum and Isaac (2016) investigate the effects of training systems on biased data in predictive policing in the U.S.. Sophisticated forecasting software is used by American police forces to construct heat-maps of cities, indicating neighbourhoods with high estimated probability of violent crimes occurring. The authors find that the training data, produced by historical patrolling efforts, are not at all good random samples, nor do they accurately represent cities' crime distributions. As a result, neighbourhoods that have been historically plagued by police forces' prejudice are constantly classified as high risk, leading to disproportionate patrolling in these areas. Due to heightened presence of patrol cars in these areas, the odds of arrests are much greater here than in less heavily monitored areas, despite these odds not corresponding to actual crime rates. This vicious cycle therefore places these historically black or Hispanic neighbourhoods at an unfair disad-

vantage, and makes it increasingly difficult to alleviate bias in criminal justice. Thus, systems that blindly learn from biased examples will continue to perpetuate past prejudice. Due to data mining algorithms reliance on training data as ground truths, it is of great importance that decision-makers assess the validity of the examples used for learning.

The second manner in which training data can lead to bias occurs during data collection, and is commonly known as *sample size disparity*. This concerns under-representation of a subpopulation due to non-random sampling. Inaccurate representations of groups of individuals can still be problematic, even if the recorded examples are free of human prejudice. Namely, under-representation of subgroups leads to decreased prediction accuracy for the disadvantaged group, as opposed to the over-represented group. As a result of poor prediction accuracy for under-represented groups, typically trustworthy criteria for fairness can result in discriminatory decision-making, as discussed in Subsection 2.5.2. Under-representation typically occurs for minority groups or third-world countries, due to lower levels of technological and online integration or economic participation, leaving them at the outer regions of today's data-generating efforts (Barocas & Selbst, 2016).

2.4.3 Feature selection.

Unwanted bias can also occur from a higher level of data-related fallacies, namely when deciding which attribute to consider or not, also known as feature selection. The core of this problem lies in the fact that data are by definition reductive representations of real-world entities or processes that can be described with infinite accuracy (Barocas & Selbst, 2016). Thus, machine readable data will never fully capture the detail of real life. More so, due to restrictions imposed by computational limitations or model interpretability, it is often necessary to compactly describe statistical relationships in a small number of attributes. Barocas and Hardt (2017) argue that some features may be much less informative or less reliable recorded for minority groups in a population. Prediction accuracy is often cited as the most important motivation to include or exclude a feature. This increase in accuracy is often paired with a decrease in fairness, referred to as the fairness-accuracy trade-off (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012; Zafar, Valera, Rodriguez, & Gummadi, 2015). Accuracy, for that matter, is also an ambiguous metric when taking fairness into account. Two models with the same prediction accuracy, for instance, can have very different subgroup accuracies due to inaccurate representation of minority groups by limited features.

2.4.3.1 Proxies.

A subcategory of feature selection related causes of bias, known as proxies, has gained considerable attention in current literature, due to historical relevance (Barocas & Selbst, 2016; Corbett-Davies & Goel, 2018). These are superficially accepted characteristics that can be used as approximations for protected attributes, such as ethnicity and religion. Proxies are often found to be the cause of disparate impact. A well-known example is *Redlining* in the United States and Canada in the 20th century, i.e., denying various services or opportunities based on applicant's postal codes. More specifically, malicious decision-makers used ZIP codes as proxies for ethnicity, as certain neighbourhoods are historically inhabited by minority groups, predominantly (d'Alessandro et al., 2017).

2.5 Formally defining algorithmic fairness

Intuitively, fairness is something most humans have an innate understanding of, albeit often subject to a person's characteristics. Mathematically, however, defining fairness seems like a never-ending multi-angle tug of war, satisfying one set of criteria as it violates the next. As one can imagine,

with the ability to ethically do so, is a challenging task, to say the least. Computer scientists and statisticians have come up with a plethora of formal definitions of fairness, relying on the satisfaction of various sets of criteria. Narayanan (2018), for instance, manages to touch upon 21 formalisations of fairness. Sadly, no all-encompassing definition of fairness has been found, nor will this ever happen. It is therefore of great importance to be able to identify which formal definition of fairness applies to which use case, to ensure equitable outcome for all members of society.

2.5.1 Formal setup & terminological and notational conventions

As previously mentioned, studying fairness in machine learning applications is a relatively new practice. As is the case with most emerging fields of study, terminology and notation are not yet unanimously agreed upon, sometimes causing ambiguous interpretation of scientific writing. To make the interpretation of the remainder of this thesis easier, several notational conventions are introduced here. The following notational rules hold, unless explicitly stated otherwise: lowercase italic Roman letters are used to indicated scalar variables. Random variables are denoted by italic type uppercase Roman letters, and calligraphic uppercase Roman letters are used to indicate an unspecified, unnamed, or generalised distribution. For instance, $R \sim \mathcal{R}$ is equivalent to saying that a random variable R follows a probability distribution denoted by $\mathcal{R}$. A general realisation of a random variable is denoted by, e.g., $R = r$. Matrices and vectors are represented as boldface upper- and lowercase letters, respectively.

Furthermore, because binary classifiers and their error types are commonly studied in fair ML research, a few terminological conventions are stated here. For the types of decision-making processes that are often the subject of fairness study, a binary outcome variable, typically denoted Y, has a *favourable* ($Y = 1$) and an *unfavourable* ($Y = 0$) outcome class. The same holds for a corresponding binary prediction label, commonly referred to as $\hat{Y}$. For the sake of semantic consistency, I choose to let outcome 1 denote the *positive* (i.e., favourable), and 0 the *negative* (i.e., unfavourable) outcome classes or prediction labels. For instance, in a recidivism prediction context receiving a label $\hat{Y} = 1$ while belonging to the outcome class $Y = 0$ corresponds to being labelled as *low*-risk (i.e., the predictor expects this individual to *not* recidivate), while in fact being re-convicted of a crime in the future. This example is known as a *false positive* error. See Section 4.1 for a more detailed overview of a generalised binary classifier's performance measures.

Let $\mathbf{X}_{i,:} \in \mathbb{R}^{(1\times k)}$ denote the set of individual i's observable features, where $i = 1, ..., N$, making the total set of observed features for all individuals an $N \times k$ matrix $\mathbf{X}$. Corbett-Davies and Goel (2018) partition these observed features into *protected* and *unprotected* attributes: $\mathbf{X} = [\mathbf{X}^{(p)} : \mathbf{X}^{(u)}]$ (for an individual, this partitioning corresponds to $\mathbf{x}^T = [(\mathbf{x}^{(p)})^T, (\mathbf{x}^{(u)})^T]$), where $\mathbf{X}^{(p)}$ and $\mathbf{X}^{(u)}$ have p and u columns, respectively. Protected or sensitive attributes are lawfully defined traits, on the basis of which an individual might be discriminated or experience disparate impact. Examples include ethnicity, sex, gender identity, religion or sexual orientation. Whether an attribute is deemed protected is application specific (Bellamy et al., 2018). Namely, in one use case, discriminating on the basis of a certain attribute can be perfectly acceptable, whereas basing decisions on the same characteristic could be viewed as malicious practice in another. For instance, taking gender into account when hiring security guards in male detention centres is perfectly permissible, whilst a restaurant owner cannot do so when hiring staff.

Hardt et al. (2016) and Barocas et al. (2018), however, use a simpler indicator variable, $A \in \{0, 1\}$, to denote protected group membership, where $A = 1$ denotes an individual belonging to a sensitive class, and $A = 0$ otherwise. For instance, we might have $A = 0$ for males and $A = 1$ for females. This thesis focuses on binary prediction tasks (an applicant will be admitted to a university or not, a defendant receives a low or high recidivism risk score, etc.), but the notational conventions

introduced in this subsection can easily be extended to multiclass classification problems. For the remainder of this report, a mention of observable features $\mathbf{X}$ and sensitive attribute indicator A in the same context implies $\mathbf{X}$ consists of unprotected attributes exclusively, unless stated otherwise. Define a binary predictor, $\hat{Y} := \hat{y}(\mathbf{x}, A)$ or $\hat{Y} := \hat{y}(\mathbf{x})$ where $\hat{Y} \in \{0, 1\}$, trained on the observed data. The goal of this predictor is to approximate some target variable, Y. Note that these are all random variables in the same probability space. That is, we assume that realisations of these stochastic variables are samples from the joint distribution $(\mathbf{X}, A, Y) \sim \mathcal{J}$. To summarise, the goal of fair machine learning is to predict some true outcome Y, using a learned predictor $\hat{Y}$, based on features $\mathbf{X}$, whilst ensuring non-discrimination with respect to sensitive attribute A.

Finally, we introduce the concept of real-valued (risk) scores. Decision-making processes essentially try to approximate an individual's risk distribution and base a choice on the estimated probability of a certain event occurring. Put differently, they approximate $\Pr(Y = 0|\mathbf{X})$, that is, the conditional probability of reoffending ($Y = 0$), given observed features $\mathbf{X}$. In practice, decision-makers tend to use real-valued predictive scores $R = r(\mathbf{x}, A)$, such as FICO scores for predicting creditworthiness introduced in Subsection 2.3, or COMPAS' recidivism risk decile mentioned in the Introduction. Please note that these scores need not lie in the interval, $[0, 1]$. Typically, higher values of R should coincide with a greater estimated likelihood of $Y = 0$, and therefore a tendency to predict $\hat{Y} = 0$. When taking, for example, COMPAS scores ($\in \{1, 10\}$), a defendant with a score of $R = 4$ is expected to be less likely to recidivate than a defendant with $R = 8$.

A binary classifier can easily be obtained from a risk score by thresholding, namely, by requiring $\hat{Y} = \mathbb{I}\{R \leq \tau\}$ for some threshold, $\tau \in \text{ran}(r(X, A))$. Here, $\mathbb{I}\{...\}$ denotes an indicator function, equalling one if its argument is true and zero otherwise. For instance, users of COMPAS scores often use a threshold of $\tau = 4$ to distinguish between defendants with a low or high perceived likelihood of reoffending (Angwin et al., 2016). A benefit of opting for such a threshold approach is that the trade-off between a binary classifier's true positive rate and false positive rate can be measured and plotted by varying the threshold, yielding a receiver operating characteristic curve or ROC curve Hardt et al. (2016).

Finally, as a purely practical convenience, consider the following notational convention. Denote the probability of an event E occurring, conditional on group membership $A = a$, as $\Pr_a(E) = \Pr(E|A = a)$.

2.5.2 The three fundamental principles of algorithmic fairness

As previously mentioned, there is no scarcity of formal mathematical definitions of fairness in the current literature (Narayanan, 2018). Despite there being myriad ways of defining algorithmic fairness, Barocas et al. (2018) argue that, fundamentally, *most* of these definitions are reducible to three criteria. Their reasoning is based on the assumption that the majority of fairness criteria impose constraints on the joint distribution of the target variable Y, the risk score R or classifier $\hat{Y} = \mathbb{I}\{R \leq \tau\}$, and the protected attribute A. More specifically, they suggest to express the joint distribution of these three random variables in terms of three conditional independence statements. These three fundamental criteria are *independence*, *separation* and *sufficiency* (see Table 1).

Clarity is an obvious benefit of being able to categorise fairness definitions according to these principles. Furthermore, once one understands the advantages, drawbacks and legal repercussions of each fundamental criterion, evaluating applicability of new fairness definitions becomes more feasible, as its becomes a matter of correctly classifying the new notion of non-discrimination. Another convenience of the three principle is their capability to be depicted as *causal graphs* (also known as Bayesian networks or directed acyclic graphs (DAGs)), that is, a graphical representation of conditional dependencies between random variables. These are probabilistic graphical models used

to visualise the encoded conditional independence assumptions of a data generating process. Random variables are indicated by their relevant symbol enclosed in a circle, known as a node, and statistical dependencies are depicted by undirected lines, known as arcs, vertices, or edges.

Independence	**Separation**	**Sufficiency**
$\hat{Y} \perp A$	$\hat{Y} \perp A \mid Y$	$Y \perp A \mid \hat{Y}$

Table 1: The three fundamental fairness criteria, as proposed by Barocas et al. (2018)

2.5.2.1 Independence

This constraint simply requires that the classifier is statistically independent of the sensitive attribute. See Definition 2.1 below for a formalisation of independence.

Definition 2.1. *Independence.* The random variables $(\hat{Y}, A)$ satisfy independence if $\hat{Y} \perp A$.

This is pronounced as "Y-hat bottom A". Independence is one of the most widely used criteria for fairness and is formulated as follows. For all groups a, b and all values $\hat{y}$, we have

$$\Pr_a(\hat{Y} = \hat{y}) = \Pr_b(\hat{Y} = \hat{y}). \tag{1}$$

When $\hat{Y}$ is a binary classifier, independence is commonly referred to as *demographic* or *statistical parity* (Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015). This requires that

$$\Pr_a(\hat{Y} = \hat{y}) = \Pr_b(\hat{Y} = \hat{y}) \; \forall \; \hat{y} \in \{0, 1\} \text{ and } a, b. \tag{2}$$

This corresponds to the proportions of positive decisions being equal across all groups of individuals. In our recidivism example, independence means that the rates at which, for instance, men and women receive low and high risk scores must be equal. It is worth noting that imposing such a restriction on decision-making systems in criminal justice does not make much sense, as it is unjust to artificially alter penalisation rates just to satisfy quota. Demographic parity restrictions are more commonplace in employment applications where diversity is desired, and even then one could argue about its validity. Other variations found in both machine learning literature as well as legal doctrine are so-called p%-rules (Zafar et al., 2015). Suppose that group b is an unprivileged group, that is, has a lower probability of receiving positive outcome prediction (e.g., being hired for a job). This rule is then defined as

$$\frac{\Pr_b(\hat{Y} = 1)}{\Pr_a(\hat{Y} = 1)} \geq 1 - \epsilon, \tag{3}$$

where $\epsilon = \frac{p}{100}$. Equation 3 implies that a decision-making process is *fair* if the inequality holds. Additive constraints of p%-rules also exist, namely

$$|\Pr_a(\hat{Y} = 1) - \Pr_b(\hat{Y} = 1)| \leq \epsilon. \tag{4}$$

The most famous example of this criterion in practice is the "four-fifths" rule. In 1978 the U.S. U.S. Equal Employment Opportunity Commission (EEOC), Department of Labor, Department of Justice and the Civil Service Commission created this guideline for employee selection procedures. It states that “A selection rate for any race, sex, or ethnic group which is less than four-fifths [...] of the rate for the group with the highest rate will generally be regarded [...] as evidence of adverse impact”

(Barocas & Selbst, 2016). Hu and Chen (2018) argue that enforcement of such laws in the short run will improve the reputation of disadvantaged protected groups in the labour market in the long run.

Despite its simplicity, natural interpretation and compatibility with legal notions of fairness, independence suffers from some inconvenient shortcomings. First, it completely ignores possible *correlation* between target variable, Y. and sensitive attribute, A. In particular, this limitation rules out the perfect predictor $\hat{Y} = Y$, when the marginal distribution of the target variable is different across groups (i.e., $\Pr_a(Y = 1) \neq \Pr_b(Y = 1)$). Independence and its variants are called 'optimality incompatible', in this case. Aside from this drawback, Barocas and Hardt (2017) also warn for what they call *laziness*, namely, the erroneous practice of accepting qualified people in the advantaged group, and random (possibly unqualified) individuals in the other group, for the sake of satisfying demographic parity. This laziness can adversely affect both the decision-maker as well as the protected class. Consider a lending example, where creditworthy whites are granted loans at a certain rate p. Due to, for instance, sample size disparity (see Subsection 2.4), the prediction accuracy is much lower for blacks than for whites. However, by constraint of demographic parity, the decision-maker is required to accept loan applications from the protected group with the same rate, p. To satisfy this constraint, the decision-maker randomly accepts both creditworthy and non-creditworthy blacks. The non-creditworthy minorities will most likely fail to repay the loan, further impoverishing them and deteriorating the bank's utility. This impoverishment of minority groups will then result in a decision-making algorithm obtaining more training examples of non-creditworthy minorities, thus creating a vicious cycle that diverges from the long term goal of equal lending rates among all races. Acceptance (of unqualified), in this sense, can be a mixed blessing. In general terms, this notions allows a decision-maker to wrongly trade false negatives for false positives.

2.5.2.2 Separation

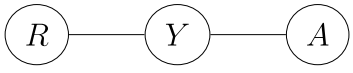
The shortcomings of independence motivated researchers to devise fairness criteria that do take the possible correlation between the sensitive attribute and target variable into account (Barocas & Hardt, 2017). Require that the score, R, and sensitive attribute, A, to be independent, conditional on target variable Y. See the following definition:

Definition 2.2. *Separation* The random variables (R, A, Y) satisfy separation if $R \perp A \mid Y$.

Formalising this as a constraint, this means that for all groups, a and b, and all values r and y,

$$\Pr_a(R = r|Y = y) = \Pr_b(R = r|Y = y). \tag{5}$$

Note that Equation 5 is essentially Equation 1, conditioned on Y. The name, separation, comes from the notion's representation as a graphical model. When we view the following diagram,



we see that the target variable nodes separates the sensitive attribute and score nodes. Intuitively, this causal graph says that the risk score R is conditionally independent of sensitive attribute A, given outcome variable Y.

Separation has the desirable property of optimality compatibility, namely, $\hat{Y} = Y$ is permissible. In other words, separation allows the perfect predictor to be a feasible solution. In particular, separation allows your target variable, Y, and sensitive attribute, A, to be correlated. Which means that correlation between A and R is also perfectly permissible. Intuitively, this makes sense, as it allows your real-valued risk score to be correlated to the protected attribute, to the extent that

is allowed by the target variable. Furthermore, separation penalises laziness as it incentivises the decision-maker to reduce errors uniformly in *all* groups, due to its requirement to have parity in both the true and false positive rates. In other words, separation equalises the cost of uncertainty across the different groups. Recall that independence satisfies neither of these desirable properties.

Hardt et al. (2016) propose two variants of separation, named equalised odds and equality of opportunity. These criteria are discussed in further detail towards the end of this subsection, and their work is covered in Section 3. They achieve separation by correcting score function R's threshold in post-processing, based on A. Note that this approach does not require retraining or applying changes to R. Given R, the trade-off between true positive and false positive rates can be plotted for all possible thresholds, yielding an ROC-curve. These ROC-curves can be constructed for both groups a and b. Visually, separation corresponds to finding the intersection of the two areas under a and b's respective curves. This intersection is known as the feasible region, encompassing all the realisable trade-offs for both groups. This is depicted by the shaded pink region in Figure 1. Now, given the application-specific costs of false negatives and false positives, the decision-maker can choose the optimal threshold in the feasible region. Note how all points in the feasible region correspond to equal false positive rates (hereafter FPR's) and false negative rates (hereafter FNR's) for both groups, thus satisfying separation's notion of non-discrimination.

Achieving separation by post-processing, though appealing, comes with some caveats. If the score function, R, is close to to the Bayes optimal score, separation via post-processing will preserve optimality among all separated scores (Barocas & Hardt, 2017; Hardt et al., 2016). However, if R is a poor approximation of the true underlying risk, the constrained separated solution will be even more unreliable. This could potentially cause harm to both the decision-maker's utility and individual well-being. If this is the case, Hardt et al. (2016) suggest to invest in collection of more reliable data and reconsider target variable labelling, or impose separation as a constraint during model training. Thus, separation implies faith in the quality of the data and predictive relevance of the target variable. Given this faith is just, separation offers desirable optimality compatibility and incentive to penalise laziness. If proven to be misplaced, data collection and model learning should be scrutinised.

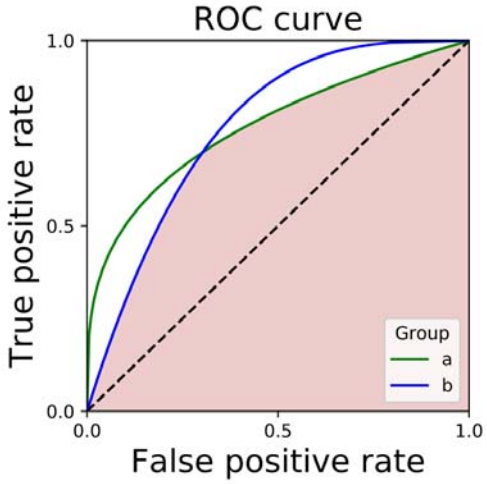


Figure 1: Visual illustration of separation for binary classifiers, with respect to two groups of individuals. Separation finds the so-called feasible region (shaded pink region) for both groups, where parity of TPR and FPR is realised. Taken from Barocas and Hardt (2017).

2.5.2.3 Sufficiency

Definition 2.3. *Sufficiency* The random variables (R, A, Y) satisfy sufficiency if $Y \perp A \mid R$.

Sufficiency, covered by Zafar, Valera, Gomez Rodriguez, and Gummadi (2017), means that the target variable is independent of the sensitive attribute, given the real-valued score. Its name, like separation, is derived from its representation as a causal model.

One can state that, for the purpose of predicting the target variable, Y, the sensitive attribute, A, becomes redundant, as the score, R, is sufficient to perform this task. Sufficiency assumes that the possible effect of A on Y is subsumed by R, and explicitly considering A is no longer needed. From a legal perspective, this is of course very appealing. In a credit scoring example, this corresponds to assuming that FICO scores already convey the possible effect of, say, gender on creditworthiness. Ruling out the need to take gender into account when extending loans.

To achieve sufficiency, Barocas and Hardt (2017) point out that it is implied by *calibration by group*. Thus, a decision-maker must calibrate his or her score function to satisfy sufficiency. Formally, this is given by

$$\Pr(Y = 1 | R = r, A = a) = \Pr_a(Y = 1 | R = r) = r, \tag{6}$$

where r is normalised to the interval $[0, 1]$. Intuitively, this means the score output r can be interpreted as a reliable probability of a positive outcome, for each group. For example, in a recidivism case using COMPAS risk decile scores, this would imply that a certain score r for both white and black defendants, corresponds to the same probability of reoffending.

2.5.3 Unawareness, individual and counterfactual fairness

It is worth noting that not all proposed fairness metrics fall neatly into the categories, independence, separation and sufficiency. Namely, some proposed methods fall into categories of their own. Some notable criteria are *unawareness*, *individual* and *counterfactual fairness*.

Unawareness (also known as anti-classification), arguably the simplest criterion for fairness, stipulates exclusion of the sensitive attribute in the training data, as discussed by Grgic-Hlaca, Zafar, Gummadi, and Weller (2016). This aligns nicely with the notion of disparate treatment, introduced in Subsection 2.2. Mathematically, unawareness amounts to assuming

$$\hat{Y} = \hat{y}(\mathbf{X}, A) = \hat{y}(\mathbf{X}). \tag{7}$$

The obvious benefit of unawareness is its simplistic intuition and applicability, and legal support for disparate treatment cases. Conversely, there may be features in X_u that are highly correlated with the protected attributes that can be used as accurate approximations for these variables (i.e., proxies). Thus, simply ignoring sensitive attributes often doesn't alleviate the problem of discrimination (Corbett-Davies & Goel, 2018).

Individual fairness, introduced by Dwork et al. (2012), differs from all previous notions of fairness, because it is, as the name suggests, focused on individuals, as opposed to group fairness. It is based on the principle that "similar individuals should be treated similarly", as stated by the authors. In order to achieve this individual-based non-discrimination, they assume a distance metric d, that quantifies the similarity between two individuals with respect to the application-specific task. Here,

$0, d(x, y) = d(y, x)$ and $d(x, x) = 0$, where x and y denote different individuals. They continue their formalisation by imposing a *Lipschitz* condition on the classifier, arguing that a classifier can be seen as a randomised mapping from individuals to probability distributions over outcomes. That is, let O denote a measurable space, and $\Delta(O)$ a measurable space of the probability distribution over O. Then, a classifier can be formalised as $M : V \to \Delta(O)$, mapping each individual to a distribution of outcomes. Two arbitrary individuals x, y with measurable distance $d(x, y)$, map to distributions of outcome $M(x)$ and $M(y)$, respectively. Now, the *Lipschitz* condition requires that the statistical distance between distributions $M(x)$ and $M(y)$ is less than or equal to $d(x, y)$, or $D(M(x), M(y)) \leq d(x, y)$ (where D is also a metric function). In the measurable probability space, this corresponds to the outcome distributions of x and y are identical up to their distance $d(x, y)$.

The benefit of individual fairness is that it devotes more attention to possible heterogeneity of the population, by imposing restrictions on each pair of individuals, as opposed to the previously mentioned group-based approaches. On the other side, a fundamental limitation of individual fairness is the assumed existence of an appropriate metric function d. Intuitively too, this hurdle cannot be ignored. How do you go about defining the similarity of two people? Kim, Reingold, and Rothblum (2018) make an argument against the method proposed by Dwork et al. (2012), by questioning the validity of assuming the existence of an appropriate metric function, and proposing an extension method, named metric multifairness. Their notion is based on the more realistic principle that "similar subpopulations are treated similarly". Despite this less fine-grained approach having more real-world interpretability, it suffers from similar shortcomings as individual fairness.

Consider the following shortcomings of all previously mentioned criteria for fairness: independence, separation and sufficiency are so-called observational fairness criteria, meaning that they convey no information regarding the potential causes of unwanted bias, unawareness is most likely too short-sighted of an approach due to its ignorance of possible correlated features, and individual fairness is limited by its necessity of an appropriate similarity measure for individuals or sub-populations. Proposed by Kusner, Loftus, Russell, and Silva (2017), counterfactual fairness, based on the principles of counterfactual inference, theoretically remedies all these issues, by providing a means to interpret the possible causes of unwanted bias. A predictor $\hat{Y}$ is said to satisfy counterfactual fairness if, given a sensitive attribute A and observable features $\mathbf{X}$,

$$\Pr(\hat{Y}_{A \leftarrow a} = \hat{y} | \mathbf{X}, A = a) = \Pr(\hat{Y}_{A \leftarrow a'} = \hat{y} | \mathbf{X}, A = a) \; \forall \; \hat{y}, a \neq a'. \tag{8}$$

Counterfactual fairness essentially requires that a decision is fair towards an individual if the outcome is the same in the actual observed world as in a counterfactual world, where the individual belonged to a different sub-population, a'. For instance, a hiring process is counterfactually fair w.r.t. an individual if it had resulted in the exact same outcome for a black applicant, had the applicant been white. The authors argue that a crucial step in alleviating unfairness, is properly addressing causality of bias, using graphical modelling. Their proposed notion does exactly that, by providing a method to check the influence of altering only the sensitive attribute. An important setback of counterfactual fairness is its infeasibility in many real-world applications. Namely, it is often difficult or even impossible to construct an effective and operational similarity measure with which to compare individuals. Furthermore, it is difficult to reach consensus on what the correct causal model should look like. This problem worsens as the number of features considered increases.

2.5.4 Examples of fairness definitions

In this subsection, I highlight a few examples of algorithmic fairness definitions, commonly used in fair ML scientific literature. Their mathematical definitions, implications, relationships to other

fairness definitions, and their applicability are briefly covered. It is important to note the difference between a fairness definition and the fundamental fairness principles mentioned earlier. The fairness *principles* are the most basic mutually exclusive requirements for (arguably) fair decision-making, of which fairness *definitions* are special cases or extensions, used in practice.

Classification parity. This is an umbrella term referring to a collection of metrics that require some classification metric (typically one derived from a confusion matrix such as false positive rates, precision and recall) to be equal across groups of individuals defined by their sensitive attributes. Classification parity falls into the independence category of criteria. Here we define demographic parity, that is equality of proportion of positive classifications, as defined by Feldman et al. (2015),

$$\Pr(\hat{Y} = 1|\mathbf{X}^{(p)}) = \Pr(\hat{Y} = 1). \tag{9}$$

Whereas parity of false positive rates is formally defined as

$$\Pr(\hat{Y} = 1|Y = 0, \mathbf{X}^{(p)}) = \Pr(\hat{Y} = 1|Y = 0). \tag{10}$$

Calibration. A member of the class of sufficient fairness criteria, calibration is a fairness measure concerning risk scores that approximate a respondent's true risk, like FICO or COMPAS scores. It requires that the risk scores, $r(X)$, correspond to the same underlying risk, independent of an individual's protected attributes, $\mathbf{X}^{(p)}$. This can be formalised as

$$\Pr(Y = 1|r(\mathbf{X}^{(u)}), \mathbf{X}^{(p)}) = \Pr(Y = 1|r(\mathbf{X}^{(u)})). \tag{11}$$

Equalised odds. Hardt et al. (2016) propose two *oblivious* fairness metrics (both falling under the separation class of fairness measures). A metric is said to be oblivious if it depends solely on the joint distribution of said metric, protected group membership and the target variable. The first proposed notion of non-discrimination is called equalised odds. The goal of this fairness criterion is to impose a non-discrimination condition, whilst aligning with the central goal of building accurate classifiers. In contrast to demographic parity, equalised odds allows the predictor, say $\hat{Y}$, to correlate with protected class membership, say A, but only through the target variable, Y. Formally, a predictor $\hat{Y}$ satisfies equalised odd with respect to protected group membership A and outcome variable Y if $\hat{Y}$ and A are independent, conditional on Y. This corresponds to

$$\Pr(\hat{Y} = 1|A = 0, Y = y) = \Pr(\hat{Y} = 1|A = 1, Y = y),\ \forall\, y \in \{0, 1\}. \tag{12}$$

This definition's alignment with the goal of high accuracy is easily shown, as $\hat{Y} = Y$ is always permissible. For instance, outcome $y = 0$ equalised odds requires that the predictor has equal false positive rates across the two groups $A = 0$ and $A = 1$, satisfying Equation 10. Similarly, when $y = 0$, the criterion enforces parity among true positive rates between the two demographics. A consequent drawback, however, is that the constraint penalises models that solely perform well on the majority group, by requiring that accuracy is equal across both demographics.

Equal opportunity. A less stringent version of equalised odds, is equal opportunity, also proposed by Hardt et al. (2016). This relaxation of the previous fairness criterion rests on the intuition that the outcome $Y = 1$ is often viewed as the "advantaged" or "privileged" outcome, like a loan application being accepted or being hired. Thus, Hardt et al. (2016) suggest to only enforce fairness in outcome within the advantaged group. In a credit scoring example this is equivalent to give people who would *not* default on a loan an equal opportunity of getting their loan application accepted. More formally, this means

$$\Pr(\hat{Y} = 1|A = 0, Y = 1) = \Pr(\hat{Y} = 1|A = 1, Y = 1). \tag{13}$$

Despite its being more lenient, equal opportunity can serve as a more relevant notion of non-discrimination in certain use cases, and generally allows for greater utility and accuracy.

2.5.5 The fairness-accuracy trade-off

A reoccurring, and in some sense unsettling, pattern in fair machine learning research is that of the reciprocity between fairness and accuracy. In short, it is typically observed that as a decision-making process becomes more equitable with respect to group-specific outcome, efficacy tends to deteriorate (Kamiran, Karim, & Zhang, 2012). To remain as general as possible, the term efficacy is sometimes used instead of accuracy, as it refers to a classifier's tendency to produce a desired result, i.e., it also encapsulates, say, precision and recall. However, they will be used interchangeably when discussing this trade-off. Similarly, fairness is used as an umbrella term, alluding to generally accepted notions of non-disparate outcome, e.g., statistical parity, equalised odds, or even individual-level fairness.

It is difficult to attribute a single cause to the frequently observed push-and-pull between fairness and accuracy. However, excluding certain mathematical incompatibilities associated with mutually exclusive fairness definitions, an often cited mechanism is the additional optimisation constraints that fairness introduce. Namely, a classifier that is solely concerned with maximising, say, predictive accuracy will search the problem space and try to find a feasible threshold to serve as a decision boundary between posterior class-membership probabilities. The obtained solution could violate a practically relevant fairness metric, such as equal false positive rates for men and women, which would be completely permissible for an accuracy-concerned classifier. Adding a fairness constraint during training can only leave the solution space unaffected in the best case scenario, but is more likely to reduce the number of permissible solutions. The same line of reasoning works the other way around, that is, starting with optimising a purely fairness-concerned classifier, and then constraining it to produce a reasonably accurate classifier will most likely lead to a less equitable classifier, as defined by the fairness constraint in question.

However, it must be noted that there is also fairness in accuracy. The aforementioned corner case of a trivial classifier that is only concerned with equitable outcome can have severely damaging effects on the individuals it decides over. Put differently, there are indirect dangers associated with basing decision-making processes solely on the satisfaction of equality quota. Think of the ways loan applicants are pushed impoverished when they are undeservedly granted credit, the societal harm caused by not incarcerating dangerous criminals, and the financial drawbacks of a company hiring someone who is unfit for a position. Fair outcome with regards to protected attributes, but decision-makers must not be driven completely by equality, and still value efficacy.

2.6 Algorithmic interventions: achieving fairness

Satisfaction of non-discrimination criteria can by achieved by three types of algorithmic interventions. The three types of techniques are based on the steps taken in a general machine learning pipeline that is concerned with building a classification model. These steps roughly correspond to, but are not limited to, collecting and processing (raw) data, learning or training a classifier, and evaluating efficacy by testing the obtained model. Therefore, the three types of fair algorithmic intervention techniques are *pre-*, *in-*, and *post-processing* (Bellamy et al., 2018). As their names suggest, pre-processing techniques adjust input data by imposing restrictions on the feature space to satisfy fairness criteria, in-processing methods impose constraints on a classifier at training time to

classifier's output to achieve non-discrimination (Barocas et al., 2018). Figure 2 provides a complete and detailed schematic overview of the various fairness intervention techniques in their corresponding locations in a model building pipeline.

These fairness processing techniques apply to all fundamental fairness criteria and their relaxations, but for the sake of intuitive interpretation, I will cover what each intervention method would entail when independence, $\hat{Y}(= \mathbb{I}\{R \leq \tau\}) \perp A$, is the desired fairness objective. Applying a pre-processing method would coincide with adjusting the input data such that it is uncorrelated with the protected attribute A (i.e., $\text{cov}(\mathbf{X}, A) = 0$). Whereas in-processing corresponds to imposing a constraint during the learning process that constructs a classifier from training data, requiring the distribution of R to be independent of A. Such a fairness constraint is imposed alongside an accuracy constraint, and can lead to conflicting optimal solutions, as mentioned in the previous subsection. And finally, achieving independence by post-processing is done by altering a learned classifier, such that the adjusted classifier is independent of the sensitive characteristic, i.e., $\tilde{Y} = \mathbb{I}\{\tilde{R} \leq \tilde{\tau}\} \perp A$. In the previous expression, tildes are used to indicate the post-processed classifier, risk score, and threshold.

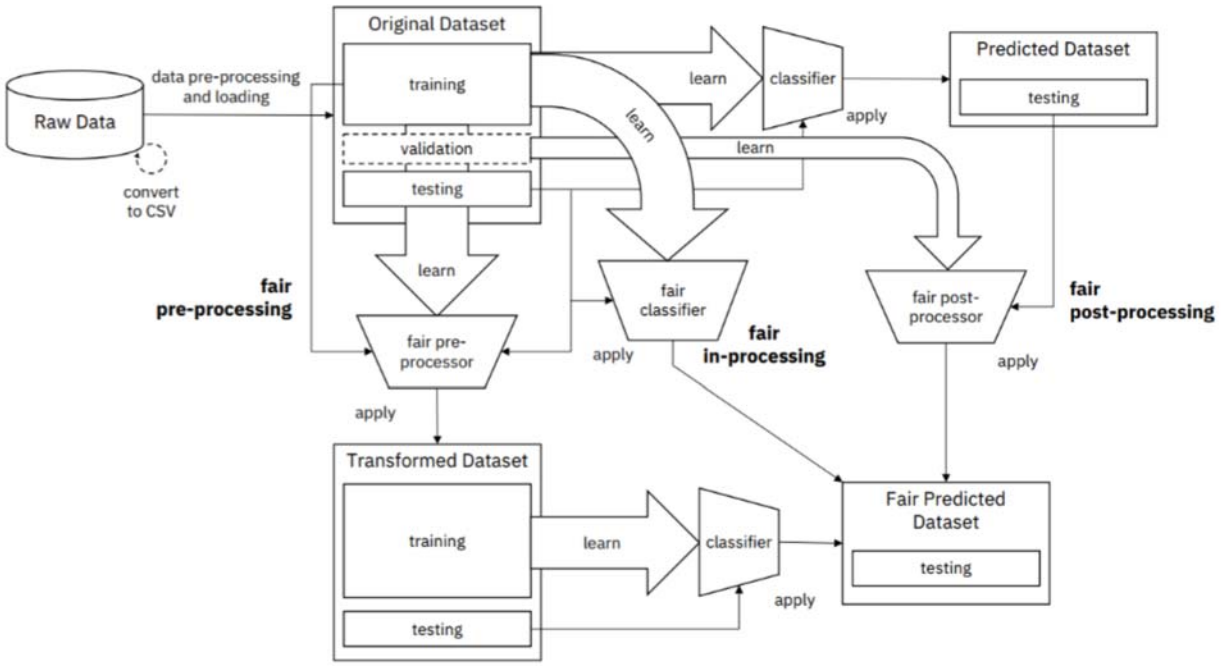


Figure 2: Schematic representation of a generalised fair machine learning pipeline. In this figure, cylinders correspond to raw, unprocessed data, rectangles are used to indicate processed data, wide arrows depict learning or optimisation processes, and trapezoids represent learned models (note that these can also be learned fairness-processors). Taken from Bellamy et al. (2018).

It is important to acknowledge the pros and cons of the three processing techniques with respect to practical feasibility and utility. Pre-processing techniques require access to the raw training data (and possibly the model building pipeline), which is not always possible in real-world applications. However, if access to raw data is possible, pre-processing techniques have the benefit of ensuring that the new and fairer feature space is unaffected by further training procedures. This follows from the information theoretic data processing inequality (DPI), which states that post-processing cannot

information content of a signal cannot be increased by any local operation. This is formalised in the following theorem.

Theorem 1 (Data processing inequality)**.** *Assume a probability model, described as a Markov Chain, formed by the three random variables X, Y and Z: $X \to Y \to Z$. That is, the conditional distribution of Z is independent of X, and depends exclusively on Y (i.e., $X \perp Z \mid Y$). The joint probability density function can be expressed as:*

$$p(x, y, z) = p(x)p(y \mid x)p(z \mid y).$$

Then it must hold that no deterministic or stochastic processing of Y can increase the information content of Y about X:

$$I(X, Y) \geq I(X, Z).$$

In the previous theorem, $I()$ denotes the mutual information of two random variables, serving as a measure of the total dependence between the two stochastic variables. For two arbitrary random variables, $I(X, Y)$ is defined as

$$I(X, Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} p_{X,Y}(x, y) \log \left(\frac{p_{X,Y}(x, y)}{p_X(x) p_Y(y)} \right) dxdy. \tag{14}$$

See Appendix C.1 for a formal proof of the DPI. The practical consequence of the DPI for pre-processing techniques is that if the dataset satisfies, say, independence, then the re-trained classifier will also satisfy independence. Note that this does not mean that the new debiased classifier will yield the same utility as its unprocessed predecessor, as it will probably be less accurate.

With preservation of efficacy in mind, in-processing is the most effective, as optimising a classifier with a fairness constraint can result in the highest possible utility of the three intervention phases (Barocas et al., 2018). However, as is the case with pre-processing techniques, enforcing non-discrimination constraints at training time requires access to the raw data, exact model specifications, and optimisation pipeline. This is unlikely to be feasible in many use cases, due to, for instance, legal restraints imposed by non-disclosure agreements. Furthermore, in-processing techniques tend to generalise poorly as they often only apply to specific model classes or training pipelines.

Finally, a post-processing technique is often the most feasible, and sometimes even the only possible option. Namely, the process of deriving a new and fairer classifier $\tilde{Y}$ from a possibly unfair and randomised classifier $\hat{Y} = f(R, A)$ that depends on a real-valued score and the sensitive attribute does not require access to the raw data on which the old classifier is trained. Moreover, post-processing techniques are applicable to any arbitrary classifier as they do not need any information on the workings of said classifier. Not having to retrain a decision-making model is a significant advantage in practice, as retraining a complex pipeline can be computationally expensive. Conversely, an obvious disadvantage of post-processing methods is the danger of severe losses in utility of a classifier (Hardt et al., 2016).

3 Related work & the COMPAS-debate

3.1 Related work

As mentioned throughout the Introduction and Subsection 2.3, data-driven criminal risk assessment practices are rapidly increasing in popularity across the globe, and even becoming well-established mandatory parts of criminal sentencing processes in a growing number of nations. These decision-making methods are also the cause of heated rhetoric between proponents and sceptics of their advent. Despite the prevalence of these algorithms and the polarising debates that surround them, there is a severe dearth of scientific evidence to back up any of the assertions made by each side of the discussion (Stevenson, 2018). There is surprisingly little empirical peer-reviewed research available about the most relevant factors criminal sentencing tools intend to influence or take into account, such as detention and crime rates, recidivism, failure to appear in court (henceforth FTA), pretrial misconduct and social disparities. Similarly, there is a lack of academic studies available about the predictive power, reliability and efficacy of these instruments. The vast majority of cited findings used to support either side of the debate comes in the form of non-academic and evidence-deficient articles or opaque and often biased reports written by the very companies that produce the risk assessment tools in question.

The peer-reviewed publications that *are* available come from various intersections of academic disciplines, like Statistics, Computer Science, Law, Criminology, Political Science and Psychology. Barocas and Selbst (2016), for instance, provide a lengthy and detailed survey of the disparities to which Big Data have given rise. They do so from a judicial standpoint, elaborating on the potential historical and technical causes of bias (discussed in Section 2.4), as well was coupling the relevant legal doctrine to their statistical counterparts. The formal legal notions of discrimination upheld in this thesis are based predominantly on their work (see Section 2.2). Their publication also serves as a good non-mathematical starting point for anyone interested in equitable decision-making in Big Data applications.

Stevenson (2018) similarly discusses algorithmic fairness from a judicial standpoint, albeit more focused on criminal risk assessment tools in particular. This article starts by pointing out that actuarial risk assessments in criminal sentencing have long surpassed the phase of being a trend and instead have become commonplace in modern courtrooms. More importantly, these instruments have attained their prominence with surprisingly little evidence-based knowledge as to their reliability and societal impact. The author then shifts the discussion from the theoretical to the practical by examining a widely used risk assessment tool used for pretrial risk estimation, the Public Safety Assessment (PSA), and its effects on racial disparities and bail amount determination in the state of Kentucky, a district heralded as a pioneer in pretrial reform. The study finds that the implementation of pretrial risk assessments failed to result in the efficiency gains anticipated by proponents, nor did it lead to racial disparities foreseen by its sceptics. The author attributes the lack of observed decreases in crime factors to three possible causes. First, the expected magnitude of decreasing rates of misconduct and other relevant factors could have been exaggerated due to the aforementioned lack of scientific evidence and understanding associated with the risk estimating instruments. Stevenson (2018) argues that the research suggesting these tools are superior to human professionals is far from conclusive. Second, it is suspected that evaluators (e.g., judges) erroneously use their authority to overrule the risk estimates when they are, in actuality, correct, instead of adjusting their own rulings when the models contradict their opinions. This possibility is attributed to an apparent widespread lack of confidence in actuarial risk assessment tools found among judicial professionals. Finally, the third suggested cause is that the risk assessment instruments did in fact improve judges' abilities

to predict future crime, but that these improvements did not result in immediate changes in the expected indicators. This failure to see improved accuracy translate into improved outcomes can be attributed to model misspecification (i.e., a failure to correctly model the ways in which crime factors are affected by certain decisions), or to the fact that risk assessments only result in improved rates of misconduct if the following actions undertaken by the evaluators are appropriate means of actually mitigating the predicted risk.

The latter suggested cause is voiced, among other points, by Berk (2017). The author argues, in accordance with the assertions made by Stevenson (2018), that debates around the use of criminal risk assessment tools consist of sparse and assumption-based rhetoric, devoid of any empirical foundation. The study examines the impact of machine learning forecasts of recidivism risk on parole release decisions made by the Pennsylvania Board of Probation and Parole. The impact of the forecasts on parole release decisions is assessed using an approximately natural randomised experiment, whereas a regression discontinuity design is used to estimate the effect on recidivism. The findings suggest that parole board members made little-to-no alternations to their choices when risk assessments were at their disposal. Inconclusive evidence suggests that the risk forecasts led to a decrease in recidivism rates. The author, however, is reluctant to draw this conclusion due to flaws in the research design, i.e. he suspects the treatment effect was inflated as a result of policy alterations when risk assessments were given to evaluators. Furthermore, the study concludes that the risk forecasts had no effect on the total parole release rate, but did change the composition of released inmates. Namely, the algorithm supposedly was able to distinguish between specific types of recidivism risks of violent and non-violent crime, thereby giving evaluators the possibility to make more nuanced (parole) release decisions.

In a more optimistic study, Kleinberg et al. (2017) report crime rate reductions up to 25% with static incarceration rates, or detention rate declines of about 42 percent without any increasing criminal rates. Using quasi-random assignment of cases to judges in a New York City dataset, and simulation studies, the authors concluded that significant efficiency gains in crime risk prediction, as well as improvements in societal disparity indicators are to be expected from the integration between machine learning forecasts and criminal sentencing. These findings, however, still fall into the aforementioned category of studies that rely on estimated results, as opposed to actually observing the hypothesised outcomes.

3.2 The COMPAS-debate timeline

In 2014, the U.S. Senate was about to pass a landmark reform bill, mandating the use of criminal risk assessment tools for sentencing processes nationwide. Eric Holder, the U.S. Attorney General at the time, voiced his concerns about how little was actually known about the ways these risk assessment instruments influence social disparities: "*Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualised and equal justice*" [...] "*they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.*" The former attorney general urged the U.S. Sentencing Commission to investigate the efficacy and long-term societal consequences of these soon-to-be mandatory risk estimation tools. The commission, however, refused to do so.

Shortly after Holder's plea goes unheard, ProPublica, an investigative journalism bureau, manages to obtain a large dataset of risk scores assigned to defendants from Broward County of Florida. Broward County courtrooms, like many other counties and states in the country, use Equivant's (then Northpointe) proprietary criminal risk assessment software, COMPAS (acronym for Correctional Offender Management Profiling for Alternative Sanctions). ProPublica was able to link these risk

of data containing a defendant's criminal record, COMPAS decile score and recidivism indicator, on which they researched the fairness and accuracy of these scores. Their findings were published in a polarising and highly publicised article that accused COMPAS scores of biased against African-American defendants (Angwin et al., 2016). The study reported, among other findings, that black defendants who didn't recidivate within a two-year period after their initial arrest were almost twice as likely as white defendants to be mislabelled as high risk criminals (false negative rates of 45 versus 23 percent for black and white defendants, respectively). False positive rates, too, were found to be skewed to the disadvantage of African-American individuals: 45% and 25% for whites and blacks, respectively. ProPublica found that even when controlling for prior offences, age and gender, black defendant's were still twice as likely to be labelled as high risk by COMPAS scores. Similar trends were reported when focusing on violent recidivism risk predictions (also provided by COMPAS).

A year later, COMPAS' developer, Equivant refutes ProPublica's findings in a report claiming that COMPAS scores are in fact fair with respect to a defendant's ethnicity, when taking into account base-rates of recidivism, a statistical aspect that ProPublica failed to control for in their investigation (Dieterich et al., 2016). The report further questions and refutes the validity of ProPublica's analyses. Dieterich et al. (2016) point out that ProPublica implemented incorrect specifications of their logistic regression model, erroneously defined classification terms and disparity metrics, and incorrectly interpreted classification errors. The authors show that when taking into account the (presumably) appropriate classification metrics, the assertions about COMPAS scores being biased towards African-Americans no longer hold. More specifically, Equivant's rebuttal questions ProPublica's use of (complements of) Sensitivity and Specificity as classification error metrics for Caucasians and African-Americans, as these measures are calculated on recidivists and non-recidivists *separately*. Dieterich et al. (2016) therefore suggest using complements of predictive values that account for base rates of recidivism among the two populations. When doing so, they conclude that COMPAS scores satisfy classification parity among blacks and whites. This means that a risk score corresponds to the same estimated probability of reoffending, irrespective of a defendant's race.

This debate between COMPAS' developer, Equivant, and ProPublica has lead to a great number of researchers joining the debate. Chouldechova (2017) and Kleinberg, Mullainathan, and Raghavan (2016), for instance, study the more theoretical side of the problem by showing the impossibility of simultaneously satisfying both claims to fairness made by ProPublica and Equivant. Similarly, Feller et al. (2016) demonstrates that both parties are essentially right and that their ideas of fairness are incompatible. Going further, they suggest that the COMPAS scores might be problematic in not yet measurable ways, i.e., increased predictive policing in neighbourhoods with high predicted recidivism rates can lead to vicious cycles of increasing racial disparities.

Barenstein (2019) examines the matter from a different point of view, by scrutinising the data manipulation choices made by ProPublica. The author accuses ProPublica of artificially inflating the recidivism rate by almost 24% by failing to apply a two-year cutoff rule for both groups (i.e., recidivists and non-recidivists) in their dataset. The study reconstructs the dataset, but corrects for the one-sided sample cutoff rule and finds that ProPublica's data pre-processing decision affects the negative and positive predictive values. However, the author concludes by noting that the choices made by ProPublica had little-to-no effect on their key statistical measures, such as false negative and false positive rates and total prediction accuracy.

4 Methodology & Data

This section outlines all the techniques, models, tools and dataset considered for this study. It lays this thesis' theoretical foundation by elaborating on the method's workings, implications and possible

mutual relationships, followed by an in depth description of the empirical data from which the results in Section 5 are derived. The current section consists of several main topics: IBM's AI Fairness 360 toolkit, quantifications of unwanted algorithmic biases in models or training data, known as *fairness metrics*, bias mitigation algorithms and the considered data.

4.1 Confusion matrices and related performance measures

A binary classifier's confusion matrix is the basis and source of many measures of efficacy and fairness in this thesis and the vast majority of fair ML literature. That is why a few frequently used confusion matrix-related notions are highlighted here for future reference in the methodology and results to come. Table 2 shows a generalised depiction of a confusion matrix. The rows correspond to the observed or true outcome label $Y \in \{0, 1\}$, and the columns reflect the classifier's predicted outcome labels $\hat{Y} \in \{0, 1\}$.

	Positive prediction class ($\hat{Y}_1$)	**Negative prediction class** ($\hat{Y}_0$)	Conditional procedure error
Positive outcome class (Y_1)	True positives tp	False negatives fn	False negative rate $\frac{fn}{tp+fn}$
Negative outcome class (Y_0)	False positives fp	True negatives tn	False positive rate $\frac{fp}{fp+tn}$
Conditional use error	Positive prediction error $\frac{fp}{fp+tp}$	Neg. prediction error $\frac{fn}{fn+tn}$	Overall error $\frac{fp+fn}{tp+fn+fp+tn}$

Table 2: Generalisation of a confusion matrix and (some of) its corresponding error type for binary classification problems. Note that the overall error is the complement of overall prediction accuracy.

A natural metric of interest is a binary classifier's overall prediction accuracy. However, this measure is inherently short-sighted, in the sense that is fails to distinguish between class-specific error types. For instance, given a highly unbalanced sample dataset where 95 out of 100 defendants are recidivists, a trivial classifier that returns the high-risk label for every input will still have an overall prediction accuracy of 0.95. To account for such shortcomings, balanced accuracy, precision, recall and F_β scores are often reported as indications of a binary classifier's efficacy. Balanced accuracy, sometimes abbreviated as $BACC$, is simply the average of the true positive and true negative rates,

$$BACC = \frac{TPR + TNR}{2}. \tag{15}$$

Precision, also referred to as positive predictive value (PPV) by Chouldechova (2017), and recall are defined as

$$\text{precision} = PPV = \frac{tp}{fp + tp} \tag{16}$$

$$\text{recall} = TPR = \frac{tp}{tp + fn}. \tag{17}$$

Finally, the F-measure or F_β score considers both precision and recall as a measure of a binary classifier's accuracy. The parameter β represents the relative weight given to recall and precision, i.e., recall is considered to be β times as important as precision. In this thesis' recidivism use-case, I have deliberately refrained from assigning more worth to either type of measure, as I believe that a proper determination of the relative importance of reducing the number falsely imprisoned defendants compared to lowering the amount of free dangerous criminals is best left to an expert in the fields of criminology, law and societal impact of either error type. A value of $\beta = 1$ is therefore chosen, making the F_1 score the harmonic mean of precision and recall:

$$F_\beta = (1 + \beta^2)\frac{\text{precision} \cdot \text{recall}}{\beta^2 \text{precision} + \text{recall}} = (1 + \beta^2)\frac{PPV \cdot TPR}{\beta^2 PPV + TPR}. \tag{18}$$

It is reasonable to want a high-stakes classifier to have equal false positive and negative error rates

equal precision across groups. However, Chouldechova (2017) shows that these criteria are not just difficult to satisfy simultaneously in practice, but in fact mathematically incompatible when risk prevalence, also known as base rate, differs among groups. This finding is formalised in Theorem 2, a proof of which is given in Appendix C.2.

Theorem 2 (Chouldechova's Incompatibility Result)**.** *If a classifier satisfies predictive parity, i.e., if PPV/precision is equal for all sub-populations/groups, but the base rates μ differ between groups, then the classifier cannot satisfy equal false positive and false negative rates across groups.*

4.2 Random Forest

Random forest, developed by Breiman (2001), is a non-parametric ensemble machine learning method for classification that creates a large collection of decorrelated decision trees using random samples from both the variable space as well as the training observations space and returns the mode of the decision trees' outputs as its classification output. Decision trees are, despite their simplicity, effective for capturing patterns in input data, and are generally low-bias (Friedman, Hastie, & Tibshirani, 2001). They are however known to suffer from high variance, and can therefore be dramatically improved by aggregating over a large number of decorrelated identically distributed decision trees. The formal process of random forest is described in Algorithm 1.

Algorithm 1 Random forest for classification

initialization

for *b = 1 to B* **do**

1. Take a bootstrapped sample of S_b size N_b from the training set.
2. Grow decorrelated decision tree T_b using S_b
   while *minimum node size n_{min} is not reached* **do**
   (a) Randomly draw m variables from the p predictors
   (b) Select the best split-point out of the m
   (c) Split parent node into two children nodes

   end
3. Return the class membership prediction of T_b, $\hat{C}_b(x)$

end

Output: Ensemble of decision trees with majority vote of output class: $\hat{C}^B_{\mathrm{RF}}(x) = \mathrm{mode}\{\hat{C}_b(x)\}_1^B$

From a fairness analyses perspective, random forest is an appealing classification method as it provides a natural way to rank the relative importance of predictors with respect to the classification task. This relative feature importance (henceforth RFI) analysis is also proposed by Breiman (2001).The two most common measures of variable importance for random forests are based on predictive accuracy and Gini impurity. Because of its more natural and intuitive interpretation regarding recidivism prediction, this study is done using accuracy-based RFI. The out of bag (abbreviated OOB) samples are used to measure prediction accuracy at the b-th tree. After accuracy is recorded, the values of a single variable j are randomly permuted in the OOB samples, and predictive strength is computed again. A decrease in accuracy after permutation is taken as a sign of feature importance. The changes in accuracy are measured and averaged over all B trees, and the resulting

feature importances are normalised, yielding a distribution of RFIs. The RFIs of all predictors are first recorded for the ensemble classifier concerned with modelling the predicted outcome label $\hat{Y}$, i.e., low- or high-risk COMPAS score, and compared to its RFI when predicting the observed outcome label Y, that is, observed recidivism. A significant difference in relative importance of protected attributes between the random forest models concerned with modelling $\hat{Y}$ and Y, respectively, can be used as tentative evidence of unfair treatment, as one would expect (protected) attributes to have similar importance scores for both $\hat{Y}$ and Y, if the decision-making model based on COMPAS scores (that implicitly defines $\hat{Y}$) is to be considered fair.

4.3 Logistic regression

To be able to draw inference about the statistical relationship between protected attributes, such as race and sex, and observed recidivism, risk score category and generalised error rates, it is necessary to make certain distributional assumptions in order to model the posterior probabilities of class membership given a defendant's observed features. Logistic regression or logit model is a natural choice for such purposes, as it provides an easily interpretable, yet generally effective manner to model the linear relationship between variables in the feature matrix $\mathbf{X}$ and the odds of belonging to a certain outcome class Y (Friedman et al., 2001).

Formally, the posterior probability of individual i with feature vector $\mathbf{x}_i$ and corresponding coefficient vector $\boldsymbol{\beta}$ belonging to outcome class $Y = 1$ is given by

$$\Pr(Y = 1 \mid \mathbf{X} = \mathbf{x}_i, \boldsymbol{\beta}) = \sigma(\boldsymbol{\beta}^T \mathbf{x}), \tag{19}$$

where $\sigma(a)$ is the cumulative density function (CDF) of the logistic distribution, also known as the logistic sigmoid function, defined as

$$\sigma(a) = \frac{\exp(a)}{1 + \exp(a)}. \tag{20}$$

4.4 IBM's AI Fairness 360 toolkit

As mentioned briefly in the introduction, IBM has responded to growing concerns regarding unfairness in machine learning by deploying an extensible architecture for understanding and mitigating unwanted algorithmic bias, named AI Fairness 360 (`https://github.com/IBM/AIF360`). The core objectives of this open source Python toolkit are to expedite the implementation of fair machine learning techniques in industrial or commercial applications and to administer an open platform for the scientific community to share and evaluate algorithms (Bellamy et al., 2018).

AIF360 consists of a comprehensive collection of datasets, fairness metrics and bias mitigation algorithms, accompanied by tutorials and the corresponding scientific publications in which the techniques are studied. These are implemented in the form of four abstractions, or so-called classes, for datasets, metrics, explainers and algorithms. The `Dataset` class and its subclasses contain tools for handling various forms of data (structured or unstructured) and built-in datasets on which models can be trained and tested. The `Metric` class and its subclasses compute various individual and group fairness metrics to quantify unwanted bias in datasets or models. The `Explainer` class works in association with the metric classes, and provides further insights about the calculated fairness measures. These explainer classes can be used to translate computed outputs to explanations that are meaningful to various industry applications. At the time of writing, IBM's AIF360 is the first fairness toolkit to emphasise the need for user-friendly and industry-specific explanations. Finally, the `Algorithms` class implements bias mitigation algorithms that aim at optimising the computed

fairness metrics by manipulating the training data, imposing constraints when learning a model or correcting outcome labels. The bias mitigation algorithms' sub-classes are categorised according to the three possible phases of intervention in a general machine learning pipeline (pre-, in- or post-processing). Currently, a total of nine bias mitigation algorithms are supported by AIF360: four pre-processing, two in-processing and three post-processing algorithms.

The following three post-processing algorithms are considered for this study: equalised odds post-processing, calibrated equalised odds post-processing and reject option classification. Equalised odds post-processing optimises equalised odds (see 2.5.4) by solving a linear program to find a new probability distribution with which to alter the outcome labels (Hardt et al., 2016). Calibrated equalised odds post-processing also attempts to optimise equalised odds, but simultaneously tries to satisfy calibration of a classifier's score output (Pleiss, Raghavan, Wu, Kleinberg, & Weinberger, 2017). Reject option classification re-assigns favourable outcome labels to unprivileged individuals and unfavourable labels to privileged individuals, within the reject option clause: a confidence region around a classifier's decision boundary with the highest level of uncertainty. (Kamiran et al., 2012). The details and workings of these algorithms are covered in Subsection 4.6.

4.5 Fairness metrics

AIF360 supports a wide variety of quantification measures of unfairness. We consider five of those fairness metrics in the current study, to provide a comprehensive view of potential unwanted algorithmic bias from multiple angles. The first four considered fairness metrics are simply computable from a binary classifier's confusion matrix. The fifth metric, known as the *generalised entropy index* (hereafter GEI), introduced by Speicher et al. (2018), unifies many prevailing measures unfairness by calculating a scalar degree of between- *and* within-group fairness.

In the following equations, we uphold the notational convenience introduced in the final paragraph of Section 2.5.1, i.e., the probability of an event E, conditional on group membership $A = a$, is subsumed in the right hand side of the following equation, $\Pr(E|A = a) = \Pr_a(E)$. Furthermore, the following formulas assume group membership $A = a$ corresponds to a majority or privileged group, and $A = b$ implies a minority or unprivileged group.

4.5.1 Statistical parity difference

This is the difference in probability of favourable outcomes between majority and minority groups (Bellamy et al., 2018). For this fairness metric, we assume that a favourable outcome corresponds to the positive prediction class (i.e., $\hat{Y} = 1$. For instance, this could be a loan approval in a credit scoring context, being hired in a recruitment example, or being labelled as low-risk by an RPI. There are, of course, applications in which $\hat{Y} = 1$ corresponds to an unfavourable outcome, such as receiving a prison sentence in a criminal justice application. Ambiguity can be avoided in such cases by simply adopting a notational convention by switching labels.

Statistical parity difference, or SPD, formally means

$$SPD = \Pr_b(\hat{Y} = 1) - \Pr_a(\hat{Y} = 1), \tag{21}$$

where $SPD \in [-1, 1]$. A value of zero denotes exactly equitable outcome between privileged and unprivileged groups, whereas values of negative and positive SPD's correspond to lesser and greater benefit for the minority group, respectively. SPD is intuitively simple and supported by the legal notion of disparate impact (See Section 2.2). However, it completely disregards equality in error rates (i.e., FPR and FNR), thereby permitting poor group-specific classifiers. As previously mentioned,

therefore consider both group-specific and all-round prediction measures when evaluating fairness through SPD.

4.5.2 Disparate impact ratio

Originally named "disparate impact" by Bellamy et al. (2018), calculates the quotient of the probability of favourable outcome between privileged and unprivileged groups. To avoid confusion with the homonymous judicial concept introduced in Subsection 2.2, this work adds the word 'ratio' to the name. The disparate impact ration (hereafter DIR) is calculated as follows.

$$DIR = \frac{\Pr_b(\hat{Y} = 1)}{\Pr_a(\hat{Y} = 1)}, \tag{22}$$

Notice that $DIR \in [0, \infty)$. Typically, there is said to be a violation of disparate impact if DIR exceeds the boundaries of some interval determined by a threshold, ϵ. For instance, a binary classifier $\hat{y}(\mathbf{X}, A)$ is in violation of disparate impact if $DIR \notin [(1-\epsilon), (1-\epsilon)^{-1}]$. $DIR = 1$ implies fair outcome of favourable decisions between both groups, whereas a value less than 1 corresponds to disadvantageous outcome for the minority group, and vice versa. DIR suffers from shortcomings, similar to those of SPD. However, it has the added benefit of being easily adaptable to existing legally defined fairness criteria, such as the four-fifths-rule.

4.5.3 Average odds difference

In contrast to the previous two fairness metrics, average odds difference (abbreviated AOD) takes into account a confusion matrix' rates of mislabelling. It is defined as the average difference in false positive rates and true positive rates between majority and minority groups, and is denoted as follows.

$$AOD = \tfrac{1}{2}\left[|FPR_b - FPR_a| + |TPR_b - TPR_a|\right] \tag{23}$$

Here, we adopt the following notational conventions: $FPR_a = \Pr(\hat{Y} = 1 | Y = 0, A = a)$ and $TPR_a = \Pr(\hat{Y} = 1 | Y = 1, A = a)$. Here, a value of zero implies equitable outcome for both groups, a negative value corresponds to disadvantageous outcome for the unprivileged group and a positive value implies higher benefit for the minority group. More specifically, notice that $AOD = 0$ implies equalised odds is satisfied (Hardt et al., 2016).

4.5.4 Equal opportunity difference

This is a special case of the average odds difference, and corresponds to the difference in true positive rates between privileged and unprivileged groups.

$$\begin{aligned} EOD &= TPR_b - TPR_a \\ &= (1 - FNR_b) - (1 - FNR_a) \\ &= FNR_a - FNR_b \end{aligned} \tag{24}$$

4.5.5 Generalised entropy index

The four fairness metrics mentioned above are all group fairness measures and suffer from common drawbacks. Namely, they do not guarantee individual-level fairness (similar individuals should be treated similarly) and fail to take into account relative group size, even though this quantity matters

fairness metrics tend to neglect societal well-being and are often infeasible due to difficulties in defining an appropriate similarity metrics with which to compare individuals (Kim et al., 2018; Speicher et al., 2018). Realising how many salient definitions of fairness tend to overlook the trade-off between individual- and group-level fairness, Speicher et al. (2018) propose to quantify algorithmic unfairness using a family of inequality indices typically used in econometrics and social sciences, known as *generalised entropy indices*. Originally proposed as a measure of income inequality in populations, generalised entropy indices or GEI's, such as the Coefficient of Variation, Gini, Atkinson or Theil index, can be interpreted as a quantification of information theoretic redundancy in data. To quantify these unfairness components, we must adopt a *benefit vector*, $\mathbf{b} = (b_1, ..., b_n)$, denoting the relative benefit of a decision-making process' outcome for all individuals in a population of size n. The family of generalised entropy indices in then defined as

$$\mathcal{E}(\alpha) = \begin{cases} \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} \left[\left(\frac{b_i}{\mu} \right)^{\alpha} - 1 \right], & \alpha \notin \{0, 1\}, \\ \frac{1}{n} \sum_{i=1}^{n} \frac{b_i}{\mu} \ln \frac{b_i}{\mu}, & \alpha = 1, \\ -\frac{1}{n} \sum_{i=1}^{n} \ln \frac{b_i}{\mu}, & \alpha = 0. \end{cases} \tag{25}$$

Where μ is the mean benefit over the entire population and $\alpha \geq 0$ is a regulation parameter which denotes the weight given to the distances between an individual's own and mean group benefit. More specifically, for large α, $\mathcal{E}(\alpha)$ is more sensitive to large deviations in relative benefit, whereas for small α, $\mathcal{E}(\alpha)$ becomes more sensitive to smaller differences in benefit. Furthermore, the benefit function is defined as $b_i = \hat{y}_i - y_i + 1$, where $\hat{y}_i$ and y_i denote individual i's predicted and true outcome labels, respectively.

Notice that several commonly used inequality indices are special cases of $\mathcal{E}(\alpha)$. For instance, $\mathcal{E}(1)$ is the Theil index, $\mathcal{E}(0)$ is the mean log deviation, and $\mathcal{E}(2)$ is $\frac{1}{2}$ times the squared coefficient of variation.

Generalised entropy indices have the desirable property, named *subgroup decomposability*. Meaning that for any partition of the total population into an exhaustive set of non-overlapping groups, subgroup decomposability guarantees that the overall measure of unfairness can be broken down into a *within-group* unfairness part (i.e., the weighted sum of inequality in benefits distributed among individuals within each group) and a *between-group* unfairness component (the weighted sum of differences in mean benefits received by each group). Subgroup decomposability enables us to take into account the trade-off between individual- and group-level unfairness. Partition the total population into $|G|$ disjoint sub-populations, where group $g \in G$ has size $|g| = n_g$ with corresponding benefit vector $\mathbf{b}^g = (b_1^g, ..., b_{n_g}^g)$ and group mean benefit μ_g. Then, we can define the mean benefit for a sub-population or group $g \in G$ as $\mu_g = \frac{1}{|g|} \sum_{i \in g} b_i$. Equation 25's first case (i.e., $\mathcal{E}(\alpha) | \alpha \notin \{0, 1\}$) can be rewritten as

$$\begin{aligned} \mathcal{E}(b_1, ..., b_n; \alpha) &= \sum_{g=1}^{|G|} \frac{n_g}{n} (\frac{\mu_g}{\mu}) \mathcal{E}(\mathbf{b}^g; \alpha) + \sum_{g=1}^{|G|} \frac{n_g}{n\alpha(\alpha-1)} \left[(\frac{\mu_g}{\mu})^{\alpha} - 1 \right] \\ &= \mathcal{E}_W(\mathbf{b}; \alpha) + \mathcal{E}_B(\mathbf{b}; \alpha). \end{aligned} \tag{26}$$

Where $\mathcal{E}_W(\mathbf{b}; \alpha)$ is the within-group unfairness component, and $\mathcal{E}_B(\mathbf{b}; \alpha)$ the between-group part, and are thereby defined by the respective terms on the right-hand side of the first equality in Equation 26. Note that the previously mentioned group-based fairness metrics solely capture the between-

group component for $|G| = 2$, whereas $\mathcal{E}(\alpha)$ takes individual-fairness into account *and* is extensible to multi-group frameworks (i.e., $|G| > 2$).

Thus, GEI's unify group- and individual-level fairness metrics into one scalar value, whilst considering the often overlooked trade-off between the two components. Furthermore, this approach provides a way to assess how unfair an algorithm is with respect to varying protected groups and intersections thereof within the same population. Finally, generalised entropy enables a decision-maker to account for *fairness gerrymandering*, that is, strategically manipulating boundaries and intersections of subpopulations in order to benefit a particular group.

4.6 Bias mitigation algorithms

Besides supporting several means of quantifying unwanted algorithmic bias, AIF360's `Algorithms` class provides a number of pre-, in-, and post-processing debiasing algorithms. These methods attempt to reduce unfair algorithmic bias by either manipulating training data, imposing constraints during model learning or altering prediction outcomes. In this study, we consider three cases of the last category, namely, post-processing algorithms. This section covers the mathematical details, theoretical implications and pseudo-code of each algorithm.

4.6.1 Equalised odds post-processing

Equalised odds post-processing, proposed by Hardt et al. (2016), optimises for equalised odds by solving a linear program. It is a supervised learning algorithm and can be executed by deriving a new predictor from an existing binary classifier or a real-valued score function, which in turn can be made a classifier by thresholding. In other words, we attempt to find an equalised odds predictor $\tilde{Y}$ derived from a real-valued score R or binary predictor $\hat{Y}$. Being the result of a post-processing procedure, we can state that the derived equalised odds predictor $\tilde{Y}$ is independent of observed features $\mathbf{X}$, conditional on the joint distribution of the score and sensitive attribute (R, A).

Achieving equalised odds must be done subject to an accuracy-preserving constraint. We define an arbitrary loss function $\ell : \{0, 1\}^2 \to \mathbb{R}$, that takes the predicted value and true label as input arguments and outputs the associated loss or inaccuracy, that is, $\ell(\hat{y}, y) \in \mathbb{R}$. The goal of deriving an equalised odds predictor is to minimise the expected loss $\mathbb{E}[\ell(\tilde{Y}, Y)]$ subject to equalised odds.

Deriving the equalised odds predictor $\tilde{Y}$ from a binary predictor $\hat{Y}$ and binary sensitive attribute A is equivalent to solving a linear program in four variables, and has a useful geometric intuition behind it. First, consider the following notational convention

$$\boldsymbol{\gamma}_a(\hat{Y}) := (\Pr(\hat{Y} = 1|A = a, Y = 0), \Pr(\hat{Y} = 1|A = a, Y = 1)). \tag{27}$$

That is, $\boldsymbol{\gamma}_a(\hat{Y})$ denotes the vector of false and true positive rates of a predictor $\hat{Y}$ within a certain sub-population $A = a$, respectively. We can now express equalised odds in terms of $\boldsymbol{\gamma}_a(\hat{Y})$. Namely, a predictor $\hat{Y}$ satisfies equalised odds *if and only if* $\boldsymbol{\gamma}_0(\hat{Y}) = \boldsymbol{\gamma}_1(\hat{Y})$, for a binary sensitive attribute.

Next, let $P_a(\hat{Y})$ denote the two-dimensional convex polytope or polygon defined as the convex hull of four vertices,

$$P_a(\hat{Y}) := \text{Conv}\Big\{(0, 0), \boldsymbol{\gamma}_a(\hat{Y}), \boldsymbol{\gamma}_a(1 - \hat{Y}), (1, 1)\Big\}. \tag{28}$$

Graphically, this can be viewed as follows. Let the newly learned predictor's false positive rates (i.e., $\Pr(\tilde{Y} = 1|A, Y = 0)$) for all possible thresholds be the x-axis, and the y-axis corresponds to $\tilde{Y}$'s true positives rates for all possible thresholds (i.e., $\Pr(\tilde{Y} = 1|A, Y = 0)$). So, $P_a(\hat{Y})$ is a closed subspace of $\mathcal{P} := [0, 1]^2$, denoted $P_a(\hat{Y}) \subseteq \mathcal{P}$. Now, $P_a(\hat{Y})$ is equivalent to the area spanned by the

lines passing through the coordinates in Equation 28, where $\boldsymbol{\gamma}_a(\hat{Y})$ and $\boldsymbol{\gamma}_a(1-\hat{Y})$ correspond to the derived predictor being equal to the original predictor $\hat{Y}$ and its complement, respectively. Notice that $P_a(\hat{Y})$ for $a \in \{0,1\}$ characterises all the *feasible* trade-offs between false and true positive rates that can be achieved by any classifier $\tilde{Y}$ derived from a *binary* predictor $\hat{Y}$, and thus, $\boldsymbol{\gamma}_a(\tilde{Y}) \in P_a(\hat{Y})$.

Finding the equalised odds solution can be summarised in the following optimisation scheme.

$$\begin{aligned} &\min_{\tilde{Y}} \mathbb{E}\Big[\ell(\tilde{Y}, Y)\Big] \\ &\text{s.t. } \forall a \in \{0,1\} : \gamma_a(\tilde{Y}) \in P_a(\hat{Y}) \\ &\gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y}) \end{aligned} \tag{29}$$

To show that Equation 29 represents a linear program in four variables, it must be shown that the objective function $\mathbb{E}\Big[\ell(\tilde{Y}, Y)\Big]$ is in fact a linear function. Consider the following expansion of the expectation of the loss function.

$$\mathbb{E}\Big[\ell(\tilde{Y}, Y)\Big] = \sum_{y,y' \in \{0,1\}} \ell(y, y') \Pr(\tilde{Y} = y', Y = y) \tag{30}$$

Where,

$$\begin{aligned} \Pr(\tilde{Y} = y', Y = y) =& \Pr(\tilde{Y} = y', Y = y \mid \tilde{Y} = \hat{Y}) \Pr(\tilde{Y} = \hat{Y}) && (31)\\ &+ \Pr(\tilde{Y} = y', Y \neq y \mid \tilde{Y} \neq \hat{Y}) \Pr(\tilde{Y} \neq \hat{Y}) && (32)\\ =& \Pr(\hat{Y} = y', Y = y) \Pr(\tilde{Y} = \hat{Y}) && (33)\\ &+ \Pr(\hat{Y} = 1 - y', Y = y) \Pr(\tilde{Y} \neq \hat{Y}). && (34) \end{aligned}$$

Continuous statistics (e.g., a real-valued score function R) convey more information about an individual observation's estimated risk than a binary variable (e.g., a predictor $\hat{Y}$). It is therefore beneficial in terms of utility to consider deriving the equalised odds predictor $\tilde{Y}$ from a score function. Furthermore, real-valued scores are common in practical applications (e.g., using FICO scores to evaluate loan applications or COMPAS scores for recidivism prediction).

For the following derivations, we will assume $R \in [0,1]$ is the normalised version of some real-valued scoring function and can thus be interpreted as an estimation of the probability of positive outcome of the target variable Y. As mentioned before, deriving a binary predictor from a score function is done by thresholding, that is $\hat{Y} = \mathbb{I}\{R > t\}$. Assuming that the underlying true risk distributions of groups defined by A are not identical, applying a single threshold to these various groups will result in a binary predictor that does *not* satisfy equalised odds, as it is unlikely that unequal risk distributions yield equal false and true positive rates. We must therefore consider using different thresholds for different subpopulations defined by A, that is $\hat{Y} = \mathbb{I}\{R > t_a\}$. Unfortunately, this is often insufficient to satisfy equalised odds (Hardt et al., 2016).

Recall from Equation 27 that equalised odds can be formulated as requiring equality of *TPR*s and *FPR*s for different values of a. Also consider that an ROC-curve is a visual representation of a binary classifier's trade-offs between true and false positive rates for all possible decision-thresholds. It is therefore useful, when attempting to achieve equalised odds, to consider the different ROC-curves determined by group membership, or so-called A-conditional ROC-curves,

$$C_a(t) := \Big(\Pr(R > t | A = a, Y = 0), \Pr(R > t | A = a, Y = 1)\Big). \tag{35}$$

Thus, we say that a score function satisfies equalised odds *if and only if* $C_a(t) = C_{a'}(t)\ \forall t, a \neq a'$. Due to the A-conditional ROC-curves being equal in this case, any threshold t will result in the equalised odds predictor. When this is not the case (i.e., ROC-curves are different), a utility maximising approach is to set different thresholds for different A-based sub-populations. Graphically, this corresponds to each A-conditional ROC-curve having its own point in the false/true-positive plane denoting the optimal threshold. However, to align with the equalised odds criterion, these thresholds must lie at the same point in the plane (which of course represents equal false and true positive rates between demographics). This is possible if there exists a point where all ROC-curves intersect. Though, in reality the curves might not have non-trivial intersections at all. Furthermore, it is also possible that such an intersection (given it exists) corresponds to an undesirable trade-off between FPRs and TPRs, regarding utility.

Hardt et al. (2016) suggest to use a randomisation approach to fill the span of all feasible derived predictors $\tilde{Y}$ and permit intersection in the false/true-positive plane. For every subpopulation a, D_a denotes the convex hull of the A-conditional ROC-curve's image, that is

$$D_a := \text{Conv}\Big\{C_a(t) : t \in [0,1]\Big\}. \tag{36}$$

Notice that D_a is the smoothed counterpart of the polytope $P_a(\hat{Y})$ in Equation 28. However, we do not consider solutions below the false/true-positive plane's main diagonal (i.e., the line connecting $(0,0)$ and $(1,1)$), as these points are worse than random guessing, and will therefore never be desirable according to any logically defined loss function.

To derive an equalised odds predictor $\tilde{Y}$ from two A-conditional ROC-curves, the algorithm chooses a point in the intersection of their respective convex hulls as a (possibly randomised) threshold predictor for each group a. Namely, any point in D_a (i.e., the convex hull of an arbitrary group a) corresponds to a feasible trade-off in true and false positive rates for a subpopulation, and hence represents a predictor $\tilde{Y}$ based on score R. Furthermore, this predictor can always be expressed as a mixture of *two* threshold predictors, as the TPR-FPR-plane is *two*-dimensional. Given that $A = a$, the derived predictor can be seen as $\tilde{Y} = \mathbb{I}\{R > T_a\}$, where $T_a \in \{\underline{t}_a, \bar{t}_a\}$ is a randomised threshold with distribution $\Pr(T_a = \underline{t}_a) = \underline{p}_a$ and $\Pr(T_a = \bar{t}_a) = 1 - \underline{p}_a = \bar{p}_a$. This means that for every subpopulation, T_a is either equal to a fixed threshold t_a or a randomised mixture of two $\underline{t}_a < \bar{t}_a$. When T_a is a mixture and $R < \underline{t})_a$ the predictor always gives a negative output $\tilde{Y} = 0$, similarly, when $R > \bar{t}_a$ the predictor is always set $\tilde{Y} = 1$ and if $R \in (\underline{t}_a, \bar{t}_a)$ the predictor is randomly set $\tilde{Y} = 1$ with probability $\underline{p}_a$.

Put differently, $\tilde{Y}$ is constructed by choosing a trade-off point in the intersection of two convex hulls (i.e., $(\gamma_0, \gamma_1) \in \cap_a D_a$), and then for each group satisfy equalised odds (i.e., $\gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y})$) using a randomised predictor $\tilde{Y} \mid (A = a) = \mathbb{I}\{R > T_a\}$.

Notice that the feasible set of FPR/TPR trade-offs of all possible equalised odds predictors is the intersection of the areas under the two groups' ROC-curves. All logically considered solutions lie on the upper-left boundary (i.e., above the main diagonal). This convex set can be seen as the ROC-curve of the equalised odd predictor $\tilde{Y}$. Also note that the equalised odds predictor creates incentive to increase prediction performance for all groups, as the equalised odds ROC-curve is point-wise minimum of each A-conditional curve, and hence represents the minimum prediction performance out of the two A-conditional predictors. Finally, finding the optimal equalised odds trade-off corresponds to solving the following optimisation problem

$$\min_{\forall a: \gamma(\tilde{Y}) \in D_a} \gamma_0(\tilde{Y})\ell(1,0) + (1 - \gamma_1(\tilde{Y}))\ell(0,1), \tag{37}$$

for a reasonable loss function ℓ. Here, we can assume $\ell(0,0) = \ell(1,1) = 0$), as any solution between

these points is undesirable.

4.6.2 Calibrated equalised odds post-processing

Next, we discuss calibrated equalised odds post-processing, as proposed by Pleiss et al. (2017). Calibration is a crucial condition for many risk assessment applications. If a decision-making algorithm's predictions do not represent properly calibrated probabilities, satisfaction of other fairness criteria, such as equalised odds, will not necessarily avoid discrimination. A lack of understanding how calibration relates to prevailing fairness definitions and shortcomings of equalised odds post-processing, as proposed by Hardt et al. (2016), are motivations for a cross-over method that achieves calibration, whilst satisfying a relaxation of equalised odds.

If we look at the number of individuals that received some (normalised) risk score p, we say that the risk assessment tool concerned with estimating the probability of belonging to the positive class $Y = 1$ is calibrated if we find $100p\%$ of the sample to actually be instances of the positive class. Intuitively, calibration means that a risk score is a reliable estimate of the true population distribution and carries semantic meaning. Although calibration is an important criterion for equity, it is not *sufficient* to ensure fairness (Corbett-Davies & Goel, 2018).

The proposed algorithm yields a calibrated relaxed equalised odds classifier by withholding predictive information from one sub-population. The setup for this post-processing approach is similar to that of equalised odds' framework, extended to allow for probabilistic classifiers. Consider a binary classification task and assume the existence of two disjoint groups, determined by a sensitive attribute $A = a$ where $a \in \{0, 1\}$. Allow these two groups to have unequal *base rates* μ_a, i.e. probabilities of belonging to the the positive class: $\mu_0 = \Pr_0(Y = 1) \neq \mu_1 = \Pr_1(Y = 1)$ where we use a shorthand notation $\Pr(Y = y \mid A = a) := \Pr_a(Y = y)$. And let $r_0(x), r_1(x) : \mathbb{R}^k \to [0, 1]$ represent *separate* risk estimators for the respective groups, where r_a outputs the probability of a given observation $\mathbf{x} : k \times 1$ belonging to the positive class (i.e., a normalised risk score), for sub-population a. In practice, however, r_0 and r_1 can be trained jointly, implying that they are the same classifier.

For ease of interpretation, define the *generalised* false positive rate of classifier r_a for group a as $g_{fp}(r_a) = \mathbb{E}_a[r_a(x)|Y = 0]$, and generalised false negative rate of the same classifier as $g_{fn}(r_a) = \mathbb{E}_a[1 - r_a(x) \mid Y = 1]$. And we call a classifier r_a *perfectly calibrated* if $\forall\, p \in [0, 1]$, $\Pr_a(Y = 1 \mid r_a(x) = p) = p$. Note that calibration of *both* estimators with respect to *both* subpopulations is crucial for non-discrimination, as failure to meet this demand can result in risk estimates carrying group-specific information (Pleiss et al., 2017).

Define in the false-positive-false-negative plane (hereafter FP-FN plane) the region of *trivial* classifiers as those that output a constant value for every input argument: $r^c(x) = c\, \forall x, c \in [0, 1]$. Furthermore, notice that the definitions of generalised error rates and calibration imply that all trivial classifiers lie on the diagonal $g_{fp}(r) + g_{fn}(r) = 1$, corresponding to random guessing. Consequently, all reasonable classifiers (i.e., those better than random) lie *below* this boundary in the FP-FN plane.

Similarly, we can characterise the set of all calibrated classifiers for both subpopulations as a linear relationship between their generalised error rates and base rates,

$$g_{fn}(r_a) = \frac{1 - \mu_a}{\mu_a} g_{fp}(r_a), \tag{38}$$

meaning that the classifier r_a lies on a line with slope $(1 - \mu_a)/\mu_a$, starting at the origin (i.e., $g_{fn}(r_a) = g_{fp}(r_a) = 0$). This lower endpoint corresponds to the perfect classifier, whereas the line's upper endpoint is its intersection with the trivial classifier diagonal, as no calibrated classifier can predict with lower accuracy than random guessing. More specifically, the upper endpoint of the line represented by Equation 38 corresponds to the trivial classifier that outputs a group's base rate μ_a.

This is the only trivial classifier that satisfies calibration, denoted as r^{μ_a}. Furthermore, Equation 38 also implies $g_{fp}(r^{\mu_a}) = \mu_a$ and $g_{fn}(r^{\mu_a}) = 1 - \mu_a$. Finally, notice that for a given base rate μ_a, the strictly better one of two calibrated classifiers lies closer to the origin on the line of all calibrated classifiers.

As alluded to earlier in this subsection, the incompatibility of calibration and equalised odds and the desirability of simultaneously satisfying both conditions serves as the main motivation for calibrated equalised odds. The impossibility of calibrated equalised odds is formalised in the following theorem (Kleinberg et al., 2016).

Theorem 3 (Kleinberg's Impossibility Result)**.** *Let r_0 and r_1 be classifiers for disjoint groups G_0 and G_1 with unequal base rates $\mu_0 \neq \mu_1$. r_0 and r_1 satisfy equalised odds and calibration if and only if r_0 and r_1 are perfect classifiers.*

The proof (for which the reader is referred to the work by Kleinberg et al. (2016) or that of Pleiss et al. (2017)) builds on the intuition that the restrictions (i.e., unequal base rates, equalised odds and calibration) define an over-constrained set of classifiers. Geometrically, we can confirm Theorem 3 by realising that equalised odds requires both classifiers to have the same coordinates in the *FPR-FNR* plane (see Section 4.6.1), whilst the unequal base rates of calibrated classifiers r_0 and r_1 stipulate that they lie on separate lines (defined by Equation 38), solely intersecting at the origin (i.e., perfect prediction). Hence, unless r_0 and r_1 have are perfect classifiers, the equalised odds constraint must be relaxed.

Begin by defining generalised a cost function c_a per subpopulation, that imposes restrictions on generalised false positives $g_{fp}(r_a)$ and negatives $g_{fn}(r_a)$

$$c_a(r_a) = \pi_a g_{fp}(r_a) + \nu_a g_{fn}(r_a), \tag{39}$$

where π_a and ν_a are group-specific non-negative constants, interpretable as the perceived cost or weight of a false positive or negative error, respectively. Furthermore, we assume that for a given base rate, *at least one* of the two constants is nonzero. In other words, the generalised cost function $c_a(r_a)$ is zero if and only if r_a is a perfect predictor (i.e., $g_{fp}(r_a) = g_{fn}(r_a) = 0$).

We can now state that calibrated classifiers r_0 and r_1 satisfy *relaxed equalised odds with calibration* if and only if $c_0(r_0) = c_1(r_1)$. Geometrically, this condition amounts to requiring both calibrated classifiers to reside on the same isoquant or level curve.

In the algorithm's description we assume, without loss of generality, that for two well-calibrated but possibly discriminatory classifiers r_0 and r_1, $c_0(r_0) \geq c_1(r_1)$. The objective is to obtain a calibrated classifier $\tilde{r}_1$ such that $c_0(r_0) = c_1(\tilde{r}_1)$. To satisfy this equal cost constraint, we withhold predictive information for a random sample of the subpopulation defined by $A = 1$. Essentially, this means we allow the classifier $\tilde{r}_1$ to return the group's base rate with a certain probability α (i.e., the output of the calibrated trivial classifier r^{μ_1}).

$$\tilde{r}_1(x) = \begin{cases} r^{\mu_1}(x) = \mu_1 & \text{with probability } \alpha \\ r_1(x) & \text{with probability } 1 - \alpha \end{cases} \tag{40}$$

Notice that Equation 40 results in the generalised cost of $\tilde{r}_1$ being a linear combination of the costs of r_1 and r^{μ_1} with interpolation parameter α: $c_1(\tilde{r}_1) = (1 - \alpha)c_1(r_1) + \alpha c_1(r^{\mu_1})$. From this last equality, we can deduce that setting the interpolation parameter equal to $\alpha = \frac{c_0(r_0) - c_1(r_1)}{c_1(r^{\mu_1}) - c_1(r_1)}$ results in satisfaction of the equal costs constraint $c_1(\tilde{r}_1) = c_0(r_0)$, whilst preserving calibration.

Algorithm 2 Satisfying calibration and a relaxed equalised odds constraint using information withholding

initialization

Input: classifiers r_0 and r_1 s.t. $c_1(r_1) \leq c_0(r_0) \leq c_1(r^{\mu_1})$ and holdout set $\mathbb{P}_{\text{valid}}$.

Output: calibrated classifiers r_0 and $\tilde{r}_1$, satisfying $c_0(r_0) = c_1(\tilde{r}_1)$.

Determine base rate μ_1 of group 1 (using $\mathbb{P}_{\text{valid}}$) to produce *trivial* classifier r^{μ_1}.

while $c_0(r_0) \neq c_1(\tilde{r}_1)$ **do**

  construct $\tilde{r}_1$ using $\alpha = \frac{c_0(r_0) - c_1(r_1)}{c_1(r^{\mu_1}) - c_1(r_1)}$

end

4.6.3 Reject option classification

The final bias mitigation algorithm considered for this study is *Reject Option based Classification*, proposed by Kamiran et al. (2012). The original paper uses the acronym ROC, but in this thesis I will refer to the method as RObC, to avoid confusion with the well-known abbreviation for the Receiving Operating Characteristic curve. This post-processing method builds on the hypothesis that discriminatory decisions are made in the vicinity of the decision-boundary, due to, what the authors suggest, is a stronger influence of bias in this area. Moreover, RObC uses an adjustable critical region around the decision boundary to re-assign outcome class labels among group instances of advantaged and deprived individuals, to reduce discrimination (analogous to affirmative action).

Some key motivations for considering RObC are a lack of requirement to modify biased data or impose constraints during training time, and extensibility with respect to multiple attribute handling or ensemble classification.

For simplicity, we limit ourselves to a two-class single classifier problem with a binary sensitive attribute indicator $A = a$ where $a \in \{0, 1\}$ ($A = 1$ corresponds to an instance belonging to the protected group). As before the target variable is defined as $Y = y$ with $y \in \{0, 1\}$ where $Y = 1$ is the favourable label. Furthermore, we assume a probabilistic outputs the posterior probability of an instance $\mathbf{x}$ belonging to the positive class $\Pr(Y = 1 \mid \mathbf{x})$. RObC attempts to reduce discrimination, defined by the authors as statistical parity difference or SPD, which corresponds to minimising $\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)$.

Typically, a trained classifier assigns an instance to the class with the greatest posterior probability (James, Witten, Hastie, & Tibshirani, 2013). However, RObC takes a different approach by categorising instances near the decision boundary as "reject options" and labels them as belonging to the positive class if the instance is a member of the sensitive group, and vice versa.

More specifically, if the classifier produces a posterior probability of an instance belonging to the positive class with high certainty (i.e., $\Pr(Y = 1 \mid \mathbf{x})$ close to 1), then the instance is labelled according to the traditional decision rule, i.e., if $\Pr(Y = 1 \mid \mathbf{x}) > \Pr(Y = 0 \mid \mathbf{x})$ then $\mathbf{x}$ is assigned to the positive class, and otherwise to the negative class. However, if the instance is classified with low certainty (i.e., $\Pr(Y = 1 \mid \mathbf{x})$ close to 0.5), the reject option-clause applies.

Adopt a reject option for instances residing in the critical region, defined as $\max[\Pr(Y = 1 \mid \mathbf{x}), 1 - \Pr(Y = 1 \mid \mathbf{x})] \leq \theta$, where $\theta \in (0.5, 1)$. As stated earlier, Kamiran et al. (2012) suggest that instances in the critical region (referred to as rejected instances or reject options) are ambiguous and most susceptible to discrimination.

Algorithm 2 summarises RObC's procedure.

Algorithm 3 Reject option based classification (RObC)

initialization

Input: A learned probabilistic classifier R that outputs the posterior probability of an instance $\mathbf{X}_{i,:}$ belonging to the positive class $Y = 1$

Output: $\{Y_i\}_{i=1}^{N}$ class labels for all instances

Reject option decision rule:

$\forall \mathbf{X}_{i,:} \in \{Z \mid Z \in \mathcal{X}, \max[\Pr(Y = 1 \mid Z), 1 - \Pr(Y = 1 \mid Z)] < \theta\}$

If $A = 1$ **then** $Y_i = 1$

If $A = 0$ **then** $Y_i = 0$

Standard decision rule:

$\forall \mathbf{X}_{i,:} \in \{Z \mid Z \in \mathcal{X}, \max[\Pr(Y = 1 \mid Z), 1 - \Pr(Y = 1 \mid Z)] \geq \theta\}$

$Y_i = \text{argmax}_{Y \in \{0,1\}}[\Pr(Y = 1 \mid X_i), \Pr(Y = 0 \mid X_i)]$

4.7 The Broward County recidivism dataset

The dataset used by Angwin et al. (2016) consists of pretrial and probation defendants in Broward County of Florida, U.S.A. (the second-most populous county in the state), who have been assessed by COMPAS' risk estimation tool between January 1st 2013 and December 31st 2014. The recidivism risk scores are based on the COMPAS survey that each defendant must fill in within a day of his or her arrest. ProPublica then links the pretrial defendants in this body of data to data about arrests up to April 1st 2016, using a defendant's reoffence record in this two-year period as the "true" outcome variable, Y (i.e., did or did not recidivate). This variable is referred to as `two_year_recid`.

For the purpose of studying the COMPAS algorithm with respect to its efficacy and fairness, the COMPAS score is converted to a prediction variable by thresholding. The score, R is a discrete variable for which holds $R \in \{1, 10\}$, where 1 corresponds to the lowest level of estimated risk and 10 to the highest. The scores are subsequently divided by COMPAS into three disjoint categories of recidivism risk, low for $R \in [1, 4]$, medium for $R \in [5, 7]$ and high $R \in [8, 10]$. According to Equivant, medium and high risk scores are much more likely to lead to increased supervision (Dieterich et al., 2016), as a low COMPAS score corresponds to a small estimated risk of reoffending in the future. Therefore, the distinction is made between low and non-low (i.e., medium and high) scores and the resulting binary variable is used as a predictor. Or formally, $\hat{Y} = \mathbb{I}\{R \leq t\} \mid_{t=4}$. Equipped with the necessary binary true outcome and prediction variables, Y and $\hat{Y}$, it is now possible to study confusion matrices, apply (most) fairness metrics and run logistic regressions, among other analyses performed in the following section. It must be noted that this particular choice of prediction variable is subjective (i.e., the distinction could also be made between high and non-high scores) and the actual extent to which a defendant's risk score influences a judicial evaluator's decision regarding e.g., probation, parole, detention and supervision is much more intricate and dependant on other factors than a single binary decision-variable. Furthermore, the "true" outcome variable `two_year_recid` is also an approximation of a defendant's actual recidivism rate, as it only takes into account a period of two years following the initial COMPAS screening, and more importantly, the target variable only accounts for *observed* recidivism. The subjectivity associated with these choices for Y and $\hat{Y}$ surely introduce some form of bias and should be kept in mind when interpreting the results presented in Section 5.

The data contain a number of sensitive and non-sensitive attributes. The most important of which is a defendant's ethnicity, denoted by the categorical variable `race`. In this dataset, there exist six categories of race: African-American (occasionally referred to as "black"), Caucasian (also known as "white"), Hispanic, Native American, Asian, and Other. See Table 16 in the Appendix for an exhaustive list of the variables contained in the raw datasets.

ProPublica applies a number of observation drops (i.e., removing individuals with incomplete or missing arrest data) and a two-year sample cutoff rule. This rule implies that two years prior to the last date on which possible reoffence could be recorded (i.e., 01/04/2014) no new defendants enter the dataset. This results in a raw dataset of size $7,214$ containing only individuals whom have been followed for *at least* two years. Furthermore, a second dataset is created for predicting *violent* recidivism. These data are a subset of the previously mentioned *general* recidivism dataset, where a distinction is made between violent and non-violent (re-)offences. Its original size counts $4,743$ observations.

Before analyses, a small number of omissions is made. Following the methodology of Angwin et al. (2016), observations with charge dates further than 30 days away from their COMPAS-registered crimes are dropped, assuming these charge data are linked to the wrong COMPAS-cases. If no COMPAS case is found, the observations are marked as `is_recid` $= -1$. This results in the final sizes of 6172 and 4020 observations for the general and violent recidivism datasets, respectively.

It must be noted that ProPublica, for unknown reasons, failed to apply this cutoff rule to all defendants in the un-processed dataset. Namely, it did allow for recidivists to enter the dataset after April 1st 2014. Barenstein (2019) covers this data processing anomaly in great detail and concludes that ProPublica thereby created an artificially high recidivism rate. He follows, however, by noting that the key statistics such as false positive and false negative rates and prediction accuracy are unaffected by this action. Nevertheless, further analyses regarding ProPublica's decision to not apply the cutoff rule to both recidivists *and* non-recidivists is beyond the scope of this research.

4.7.0.1 Demographic breakdown

Because this thesis aims to study the COMPAS algorithm's propensity for racial and gender biases, it makes sense to briefly summarise the data in terms of individual's ethnicity and sex (see Tables 3 and 4). Dissecting the general and violent recidivism datasets with respect to race and gender, it becomes clear that African-Americans and Caucasians are both well-represented, whereas the remaining ethnic groups aren't. African-Americans make up roughly 51% of the general recidivism dataset, whereas Caucasians comprise about 34% of the observations. For the violent recidivism data, this is more balanced: approximately 48 and 36 percent of defendants who had been initially registered by COMPAS for violent offences are African-Americans and Caucasians, respectively. Conversely, the remaining races combined (i.e., Asian, Native American, Hispanic and Other) only account for roughly 15% and 16% of the observations in the general and violent recidivism datasets, respectively.

As for the gender-based breakdown, approximately 20% of observations in both the general and violent recidivism datasets correspond to females. This is in line with expectations, as women are generally underrepresented in crime and incarceration rates.

A visual representation of race-wise and gender-based breakdowns can be seen in Figure 3. The similarities between the nearly identical doughnut charts shows that racial and gender ratios are roughly equal when subsetting general recidivism into violent recidivism.

Race/ Sex	African-American	Asian	Caucasian	Hispanic	Native American	Other	All
Female	549	2	482	82	2	58	1175 (**19%**)
Male	2626	29	1621	427	9	285	4997 (**81%**)
All	3175	31	2103	509	11	343	6172
%	**51.4%**	**0.5%**	**34.1%**	**8.2%**	**0.2%**	**5.6%**	

Table 3: Racial and gender-based breakdown of the *general* recidivism dataset.

Race/ Sex	African-American	Asian	Caucasian	Hispanic	Native American	Other	All
Female	393	1	336	61	0	50	841 (**20.9%**)
Male	1525	25	1123	294	7	205	3179 (**79.1%**)
All	1918	26	1459	355	7	255	4020
%	**47.7%**	**0.6%**	**36.4%**	**8.8%**	**0.2%**	**6.3%**	

Table 4: Racial and gender-based breakdown of the *violent* recidivism dataset.

(a) General recidivism.

(b) Violent recidivism.

Figure 3: Doughnut charts visualisations of the demographic breakdowns of the general and violent recidivism datasets in Tables 3 and 4, respectively. In the inner charts, purple corresponds to Females and pink to Males. The explicit percentages have been omitted to avoid clutter.

5 Results

This section presents a detailed report and interpretation of the most important results obtained during this research. The section comprises two main parts: the first part of the analyses is done using, for lack of a better term, 'general purpose' econometric and statistical methods. The workings of these methods are discussed in Sections 4.2 (random forest) and 4.3 (logistic regression). The second part discusses results gathered by applying methods deliberately designed to assess and correct levels of unfairness associated with a decision-guiding model or algorithm.

More specifically, the first part of the results examines claims of unfairness with respect to ethnicity and sex via distributional analyses of the COMPAS risk prediction instrument, various logistic regression and random forest models. These last two methods have the benefit of being easily interpretable, yet effective for non-linear classification tasks. Finally, group-based fairness metrics and optimised generalised entropy indices are applied to assess levels of disparity, which are then corrected by bias mitigation algorithms, whilst preserving predictive utility.

5.1 Analyses I: traditional econometric methods

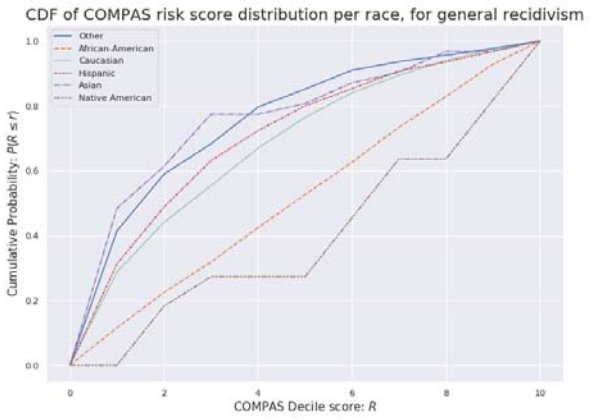
5.1.1 Distributions of COMPAS risk scores by sub-populations

To examine the distributional properties of COMPAS scores with respect to race and gender, the Cumulative Distribution Functions (CDFs) per ethnic group and sex are studied (see Figure 4). The CDF of a random variable X is defined for any real-valued x as $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$ (Bain & Engelhardt, 1987). The CDFs per ethnic group, depicted in Figure 4a, show differing distribution shapes of COMPAS risk scores among races. Specifically, the CDF of the risk score distribution among African-American (corresponding to the orange dashed line) defendants resembles that of an approximately uniformly distributed random variable, suggesting an even distribution of all COMPAS score levels. This can be seen in an alternative visual representation in Figure 5b. Conversely, the CDF of decile scores associated with Caucasian defendants (i.e., the green dotted line in Figure 4a) follows a heavily right-skewed distribution, which can also be viewed in Figure 5a. That is, the highest density of COMPAS scores of Caucasian defendants is located towards the lower risk scores. This discrepancy in risk score distribution could be interpreted as evidence of unfair treatment. However, this assertion of unfairness would only hold if African-Americans and Caucasians would have (approximately) identical base rates of recidivism. Put differently, recidivism prediction instruments, $R = r(X, A)$, aim to approximate the true underlying recidivism distribution $Y \mid X, A$. Therefore, largely differing distributions of risk scores between two sub-populations defined by protected attributes (i.e., $R(X, A = a) \neq R(X, A = b)$) may suggest inequitable outcome if the *true* risk distributions are similar (i.e., $Y_a \mid X \approx Y_b \mid X$), where, e.g., Y_a is shorthand notation for $Y \mid A = a$, where $a \in \{0, 1\}$.

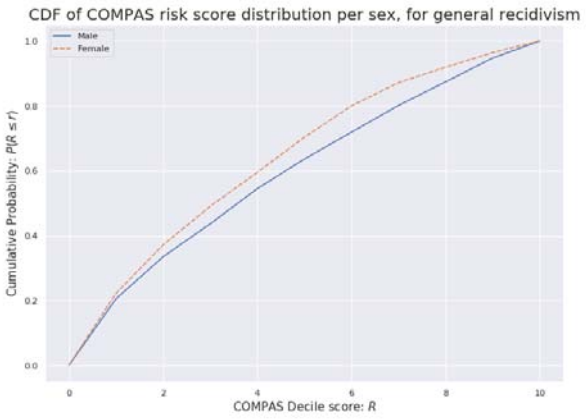
Due to the limitations of the data and the practical difficulty of obtaining reasonable estimates of an individual's true level of risk, effectively comparing the observed risk level distributions between races is beyond the scope of this research. However, comparing observed base rates of recidivism, $\hat{\mu}_a = \frac{1}{N_a} \sum_{n \in N_a} \mathbb{I}\{y_n = 0\}$, can also lead to insights with respect to distributional properties of recidivism risk among groups. Namely, as shown in Figure 6, the observed likelihood of recidivism is significantly greater for African-Americans than that of Caucasians. This average in means of the race-wise recidivism distributions implies the true underlying risk distributions of African-Americans and Caucasians are likely unequal, thereby making the discrepancy in CDF shapes between the two groups insufficient evidence for assertions of disparities, as a sub-population that shows more observed recidivism can be expected to have a greater proportion of high risk scores.

When focusing on gender-wise distributions of COMPAS risk scores, Figure 4b suggests the distribution of risk scores given to males and females are roughly equal in shape, with that of women being slightly more right-skewed (as one would expect). However, the significant difference in base rates, depicted in Figure 6b, suggests that women (who appear to recidivate with lower frequency) are being held to a similar standard as men in terms of risk assessments.

The remaining ethnic groups, i.e., Other, Hispanic, Asian and Native American, are relatively under-represented in this specific dataset, when compared to Caucasians and African-Americans. Especially Asians and Native Americans comprise such a small portion of the dataset (0.5% and 0.2%, respectively), that inferring distributional characteristics based on sample statistics is less reliable than for the two majority sub-populations (e.g., notice the large confidence regions for Asian and Native American base rates in Figure 6a). This is also the reason that these remaining ethnic groups are treated as an entire group, denoted by the dummy variable `race_other2`, in the logistic regression and random forest models, presented later in this section.

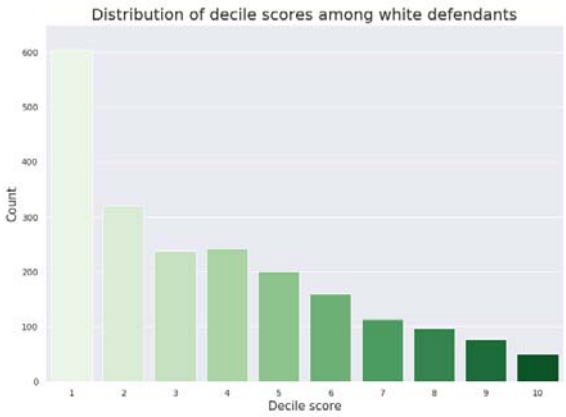


(a) CDFs per race.

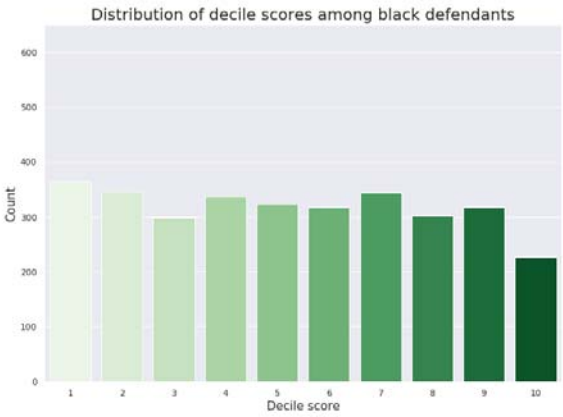


(b) CDFs per gender.

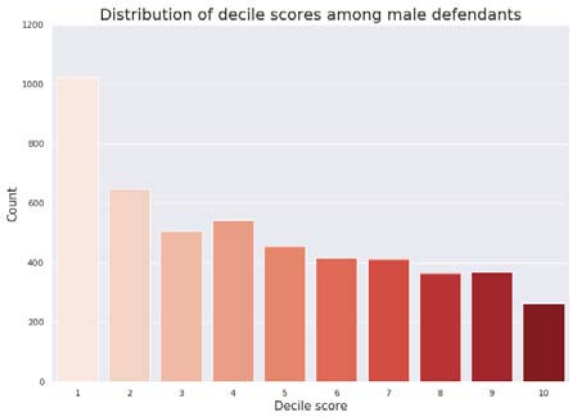
Figure 4: Cumulative distribution functions (CDFs) $P(R \leq r)$ per race, where R is the random variable generalisation of the COMPAS decile risk score.



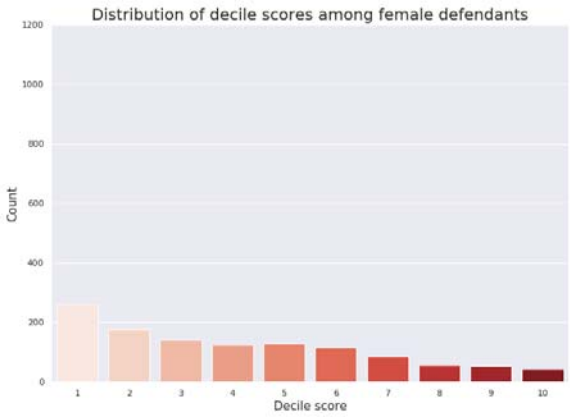
(a) Caucasians.



(b) African-Americans.



(c) Males.

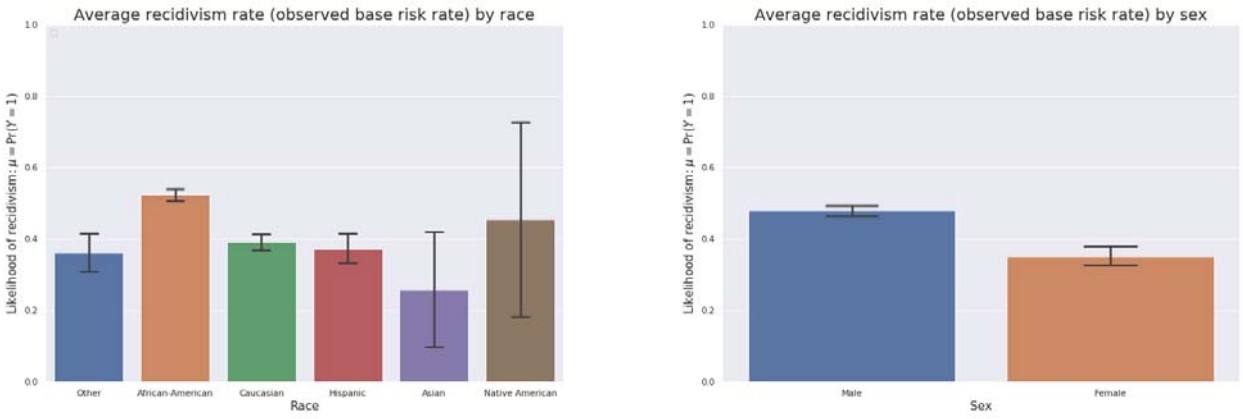


(d) Females.

Figure 5: Histograms displaying the COMPAS decile score counts by protected attribute value (i.e., race being either African-American or Caucasian, and gender being either male or female.)

5.1.2 Calibration

Intuitively, calibration of a recidivism prediction instrument (RPI) means that the probability of risk estimate outputs of this instrument are reliable estimates of the population risk distribution, for all



(a) Observed base rates per race.

(b) Observed base rates per sex.

Figure 6: Observed average recidivism rates per race (left) and sex (right). The x-axes correspond to the various sub-populations, and the y-axes display the estimated base rates of recidivism, with 95% confidence intervals plotted around the average values.

sub-populations. Phrased differently, a well-calibrated score-based classifier outputs (approximately) equal likelihood of recidivism *per risk score*, regardless of a defendant's protected attributes (e.g., race or gender). Formally, a normalised risk score r is perfectly calibrated if and only if $\forall p \in [0, 1], \Pr(Y = 0 \mid r = p, A = a) = p$, for all sub-populations $a \in |\mathcal{A}|$. For approximate calibration, the last equality symbol should be replaced by an approximation symbol. It is worth repeating that ProPublica's report has been met with much criticism, partially due to their failure to take calibration into account (Chouldechova, 2017).

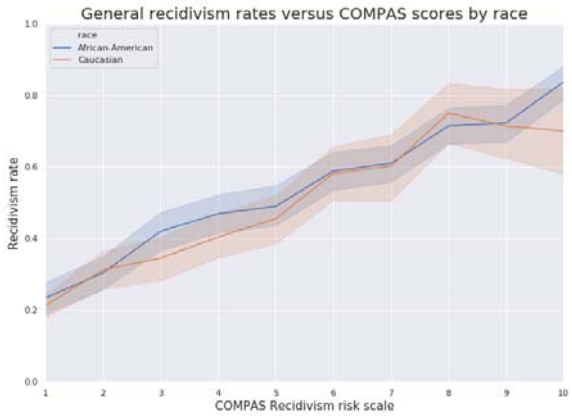
Figure 7 shows the results of testing for calibration of COMPAS scores with respect to an individual's ethnicity and gender, for general and violent recidivism. The lines represent empirical estimates of $\Pr(Y = 0 \mid R = r, A = a)$ with 95% confidence intervals depicted by the transparent bands surrounding the lines. In this visualisation, overlapping confidence intervals at a certain risk score indicate an insignificant statistical difference in estimated likelihood of recidivism between two groups at a 5% level (i.e., the RPI is well-calibrated by group at this risk score). The classifier is considered well-calibrated if and only if *all* risk score levels show no statistically significant differences in estimation of recidivism rate, that is, one non-overlapping region is sufficient to deem the classifier uncalibrated with respect to a certain protected attribute.

The COMPAS RPI satisfies calibration. This is shown in Figure 7a, where the 95%-confidence intervals overlap at all risk score levels. Similarly, satisfaction of calibration persists for violent recidivism, as depicted by Figure 7c. Conversely, COMPAS risk scores appear to violate calibration with respect to gender for both general and violent recidivism (see Figures 7b and 7d). Namely, both calibration plots show at least one pair non-overlapping confidence intervals (two for general and three for violent recidivism). The figures suggest that females are less likely to recidivate than men, despite their COMPAS scores being equal. Assuming law-enforcers apply equal decision-making thresholds to men and women, it appears that a female defendant's risk of recidivism is more likely to be unfairly over-estimated than that of a male defendant.

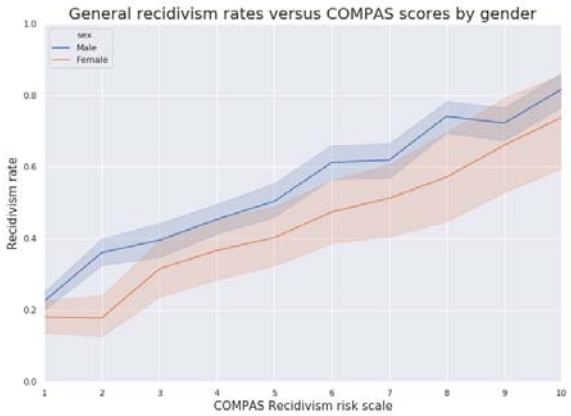
Notably, a similar RPI, the Post Conviction Risk Assessment (PCRA) tool, developed by the Administrative Office of the United States Courts for the improvement of post-conviction supervision efficacy has been shown to also be calibrated with respect to race, but not to gender (Chouldechova, 2017).

Furthermore, there is a clear difference in the width of the confidence intervals associated with the

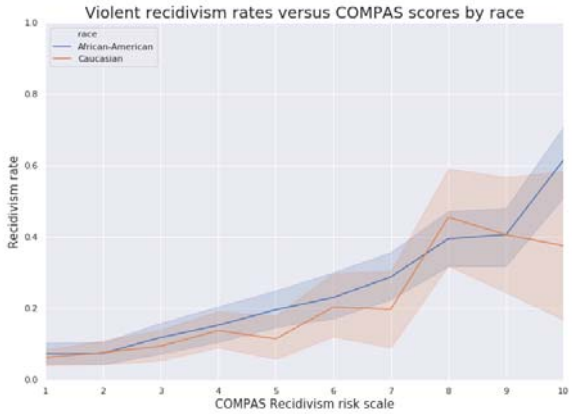
minority classes (i.e., African-Americans and women) visible in Figure 7. This holds especially for the higher parts of the COMPAS risk scale, most notably for violent recidivism. The larger confidence regions indicate greater levels of uncertainty associated with the estimates that the risk score provide, meaning that the risk estimates are less precise (i.e., they show larger dispursion around the mean) for women and African-Americans.



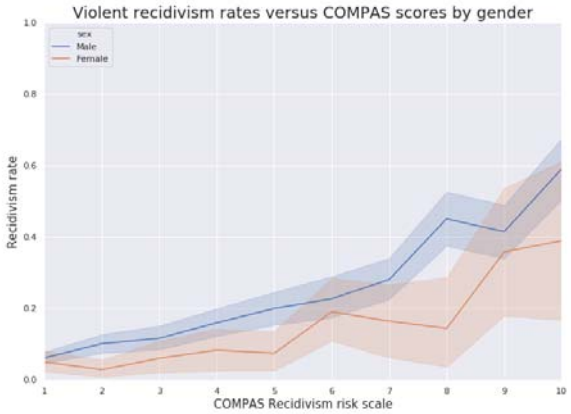
(a) Calibration line plots for general recidivism for African-Americans (blue) and Caucasians (orange).



(b) Calibration line plots for general recidivism for males (blue) and females (orange).



(c) Calibration line plots for violent recidivism for African-Americans (blue) and Caucasians (orange).



(d) Calibration line plots for violent recidivism for males (blue) and females (orange).

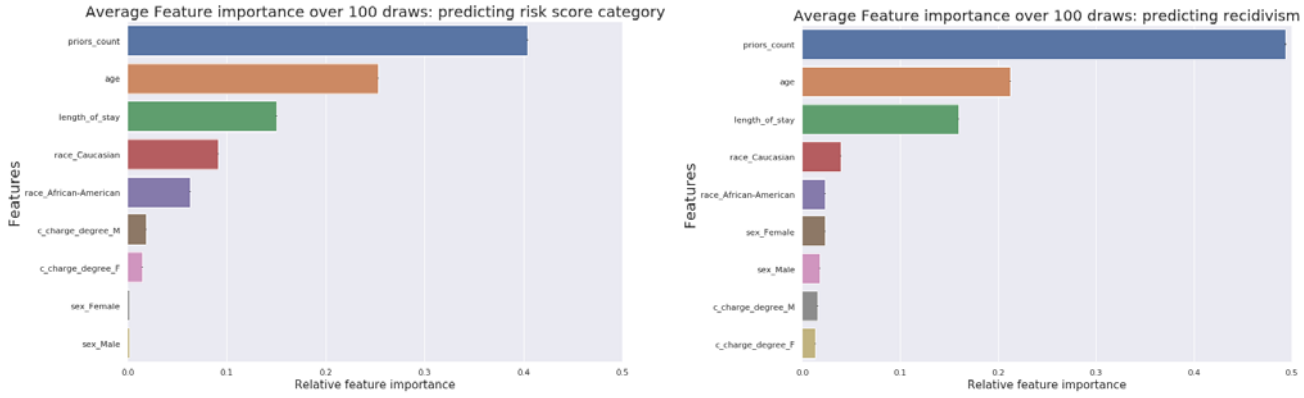
Figure 7: So-called calibration plots: line plots of (general or violent) recidivism rates per COMPAS recidivism risk score by protected attribute (race or gender). The lines represent empirical estimates of $\Pr(Y = 0 \mid R = r, A = a)$ with 95% confidence intervals depicted by the transparent bands surrounding the lines.

5.1.3 Assessing relative feature importance using random forests

In this subsection, a brief overview is presented of the most important results concerning the application of random forest models to assess relative feature importance of variables involved with the prediction of score category (low/high) and observed recidivism (no/yes). Random forests use bootstrapped samples from the training data and subsets of the feature space to grow decorrelated decision trees, aggregate them and take the mode of the output classes as final output. This is an appealing approach as it reduces variability induced by randomly growing decision trees, whilst being capable of learning highly non-linear decision boundaries. Additionally, this modelling technique is free from distributional assumptions about the data generating process. Finally, random forests provide a natural and easily interpretable framework for evaluating the relative importance of variables. These aspects make random forests a suitable modelling technique for fairness analysis, as it makes it possible to compare the relative feature importance of protected attributes when modelling the predictor variable (e.g., COMPAS score category) versus the true outcome variable (e.g., observed recidivism), while keeping performance measures in mind.

In this setup, one random forests ensemble is concerned with predicting the binary variable for COMPAS score category, `score_cat`, and the other model predicts observed recidivism within two years after the initial COMPAS screening date, `two_year_recid` (analogous to the distinction made between models in Section 5.1.4). Namely, COMPAS risk score category is defined by thresholding the decile score at $R = 4$, where scores strictly above 4 are considered high-risk (denoted by a zero), and low-risk otherwise (denoted one).

The data is randomly split into a training and test set, with respective proportions 70% and 30%. During training, a total of 100 decorrelated decision trees are grown per simulation run with maximum depth of 3. The total number of draws is 100. The relative feature importance scores are then computed during training, and presented Figures 8a and 8b for COMPAS score category and recidivism prediction, respectively. See Table 5 for the explicit numerical values of the relative feature importance scores. To assess the efficacy of the models, their overall prediction accuracy, F1-score, precision and recall are calculated and compared in Table 6.



(a) Relative feature importance histogram of random forests for prediction of score category, for general recidivism.

(b) Relative feature importance histogram of random forests for prediction of recidivism within two years, for general recidivism.

Figure 8: Histograms of relative feature importance distributions associated with two random forest models concerned with predicting decile risk score category, i.e., low or not low (left) and observed recidivism within two years (right).

As one might expect, both histograms in Figure 8 show that both random forests rank a defendant's number of prior convictions, age and length of detention period (in days) as the top three

Target variable:	**Score category**	**Observed recidivism**
Feature name	*RFI*	*RFI*
priors_count	0.404304	0.494003
age	0.253111	0.212477
length_of_stay	0.150235	0.159498
race_Caucasian	0.091286	0.039834
race_African-American	0.063213	0.023449
c_charge_degree_M	0.018507	0.015398
c_charge_degree_F	0.014976	0.013771
sex_Female	0.002318	0.023345
sex_Male	0.002051	0.018224

Table 5: Relative feature importance (RFI) scores.

	Accuracy	F_1	**Negative predictive value**	**True negative rate**
Score category	0.741	0.717	0.752	0.686
Recidivism	0.670	0.622	0.675	0.578

Table 6: Performance measures for random forests, averaged over 100 simulation runs. The entries leftmost column correspond to the dependent variable of the random forests.

both cases. A notable result, however, is the drop in relative feature importance scores of the race indicator variable (i.e., `race_caucasian` and `race_african_american`) when shifting from predicting score category (a label given by the COMPAS algorithm) and actual observed recidivism. Namely, the features indicating a defendant's race are estimated to be almost 2.5 times more important relative to the other features when predicting a defendant's score category than they are when actual recidivism is being predicted. This ratio is slightly larger when focusing on the difference in relative feature importance of `race_african_american` between score category and recidivism prediction (about 2.7).

These results must be interpreted tentatively, though. Namely, as shown in Table 6 the model concerned with predicting recidivism displays considerably worse test performance measures on average than the random forests predicting COMPAS score category, making its feature importance scores less reliable and generalisable. The true negative rate in particular, is notably inferior for the recidivism predicting model, indicating that it is more likely to incorrectly label a dangerous criminal as low-risk. Furthermore, the relative feature importance scores provide limited information as they tell us nothing about the sign or direction of the effects associated with the features, that is, the histograms in Figure 8 do not clarify whether being African-American is associated with a positive or negative statistical relationship with score category or recidivism. To obtain more directional information about the features, coefficients and thus a parametric method is required.

5.1.4 Logistic regression of score category and observed recidivism

Logistic regression models are used in this study as a parametric modelling approach to assess the statistical relationship between observed recidivism, COMPAS risk score category (i.e., low or not-low) and the protected attributes (controlled for the relevant non-sensitive covariates). From a fairness detection perspective, the purpose of these regression analyses is to study whether statistically significant relationships between the protected attributes and the outcome variables persist when modelling (predicted) COMPAS score categories versus actual (observed) recidivism.

The logistic regression analyses are set up as follows: first, the COMPAS risk score category is regressed on the two sensitive attributes and non-sensitive relevant covariates (i.e., age category, number of prior convictions, degree of offence and recidivism within two years) for both general and violent recidivism, the results of which are shown in Tables 7 and 17, respectively.

In accordance with the work of Angwin et al. (2016) and COMPAS guidelines, a threshold risk score of 4 is used to construct a binary dependent variable from a COMPAS decile risk score, that is, $\hat{Y} = \mathbb{I}\{R \leq 4\}$ (i.e., a defendant is labelled as *high*-risk if his or her COMPAS score is *greater* than 4). Put differently, the dependent variable for the logistic regressions modelling score category is expressed as

$$\hat{y} = \begin{cases} 0, & R > 4 \\ 1, & R \leq 4. \end{cases} \tag{41}$$

Then, the "true" outcome variable, i.e. observed recidivism within two years, is regressed on the same dependent variables as score category regressions, for general and violent recidivism. The results of which can be seen in Tables 8 and 18, respectively. In this setting, the outcome variable is defined as $y_i = \mathbb{I}\{\text{Defendant } i \text{ is NOT convicted of another crime within two years}\}$.

In this framework, statistically significant positive coefficient estimates indicate that the corresponding covariate is associated with an increased probability of the outcome variable being equal to zero. Intuitively, this corresponds to being labelled as high-risk (Tables 7 and 17) or reoffending within a two year span (Tables 8 and 18). Furthermore, for a given statistically significant coefficient estimate $\hat{\beta}_k$, the associated isolated change in odds of the outcome variable being equal to one, otherwise known as the estimated marginal effect of $\boldsymbol{x}_k$ on y, is equal to $e^{\hat{\beta}_k}$, *ceteris paribus* (henceforth c.p.).

The first comparison is made for general recidivism, between Tables 7 and 8. With regard to the signs and magnitudes of the coefficient estimates in the score category regression model, almost everything is in line with expectations, with the exception of the positive coefficient estimate of sex_female. The model suggests that being a woman corresponds to being $e^{0.2193} \approx 1.25$ times more likely to be labelled as high-risk, holding all other variables constant. Conversely, the recidivism regression model suggests that not being a man is associated with being $e^{-0.3477} \approx 0.71$ times as likely to recidivate within two years, c.p.. A significant positive marginal effect on being categorised as a high-risk criminal, despite showing a significant negative effect on actual observed recidivism can be interpreted as evidence of an unfair bias with respect to women. Section 5.1.2 displays a similar phenomenon, by showing that COMPAS risk scores are poorly calibrated to the disadvantage of women, even though females show a significantly lower prevalence of recidivism. The results presented in this section echo that notion, but more convincingly so, as the regression models are adjusted for other non-sensitive criminological covariates.

When focusing on the effects of race on COMPAS score category compared to those on observed recidivism within two years, the logistic regression results suggest that ethnicity is associated with significant changes in odds of being labelled as high-risk, despite not having a significant marginal effect on the likeliness of rearrest. The results in Table 7 imply that African-American defendants are 1.61 times as likely to receive high-risk COMPAS scores as Caucasian defendants, holding all other covariates equal. Being a member of the remaining ethnic groups (i.e., variable name race_other2) also displays a significant, though negative, effect on the probability of being labelled as high-risk, compared to Caucasians. However, when modelling observed rearrests within two years after the initial COMPAS-screening date, being an African-American is associated with the only insignificant coefficient estimate (race_other2 is barely significant). This could serve as evidence that ethnicity plays an important role in COMPAS score classification, but not as much in recidivism prevalence

when conditioning on relevant, non-sensitive attributes.

Similar effects are observed for violent recidivism, presented in Appendix D. Namely, the marginal effects of sex_female flip signs from positive to negative when shifting from using score category as dependent variable to modelling observed recidivism. The marginal effect of race_african_american, however, remains significant, though more moderate, when comparing score category inference to that of two year recidivism.

	coef	**std err**	**z**	***p*-value**	**[0.025**	**0.975]**
const	-1.5248	0.078	-19.442	0.000	-1.679	-1.371
sex_female	0.2193	0.079	2.764	0.006	0.064	0.375
age_cat_greater_than_45	-1.3574	0.099	-13.711	0.000	-1.551	-1.163
age_cat_less_than_25	1.3063	0.076	17.231	0.000	1.158	1.455
race_african_american	0.4770	0.069	6.879	0.000	0.341	0.613
race_other2	-0.5410	0.105	-5.165	0.000	-0.746	-0.336
priors_count	0.2695	0.011	24.305	0.000	0.248	0.291
c_charge_degree_m	-0.3089	0.066	-4.646	0.000	-0.439	-0.179
two_year_recid	0.6821	0.064	10.671	0.000	0.557	0.807

Table 7: Model specification and parameter summary of logistic regression with score category (low or not low) as dependent variable for general recidivism. Note that the variable race_other2 corresponds to a grouped dummy of all non-white and non-black ethnic groups.

	coef	**std err**	**z**	***p*-value**	**[0.025**	**0.975]**
const	-0.6082	0.065	-9.430	0.000	-0.735	-0.482
sex_female	-0.3477	0.072	-4.840	0.000	-0.489	-0.207
age_cat_greater_than_45	-0.6695	0.076	-8.801	0.000	-0.819	-0.520
age_cat_less_than_25	0.7333	0.069	10.639	0.000	0.598	0.868
race_african_american	0.0959	0.063	1.529	0.126	-0.027	0.219
race_other2	-0.1780	0.088	-2.025	0.043	-0.350	-0.006
priors_count	0.1656	0.008	20.536	0.000	0.150	0.181
c_charge_degree_m	-0.2186	0.059	-3.721	0.000	-0.334	-0.103

Table 8: Model specification and parameter summary of logistic regression with observed recidivism within two years as dependent variable for general recidivism. Note that the variable race_other2 corresponds to a grouped dummy of all non-white and non-black ethnic groups.

5.1.5 Logistic regression of false positives and false negatives

Two of the main accusations made by Angwin et al. (2016) regarding the presumed unfair racial bias are that African-American defendants are more likely to be incorrectly labelled as high-risk (false negative error) and less likely to be misclassified as low-risk (false positive error) than Caucasian defendants. These claims are based on race-wise confusion matrices derived from testing COMPAS scores as classifiers when taking $R = 4$ as a threshold for predicted risk recidivism, and comparing these predictions to the observed recidivism outcome variable. A limitation of this approach is that it does not adjust for possibly relevant crime-related variables present in the Broward County dataset. Therefore, it makes sense to reformulate their analyses in terms of logistic regression models to study whether the claims made by Angwin et al. (2016) hold when accounting for other relevant covariates.

Chouldechova (2017) applies a similar approach to assessing the statistical relationships between criminological variables and false positive errors in recidivism prediction via logistic regression, but does not extend to false negatives, and controls for a smaller number of relevant covariates.

The set up of these analyses is analogous to that of those presented in Section 5.1.4: false positive and false negative errors are taken to be the dependent variables for logistic regression models, presented in Tables 9 and 10, respectively. Of course, the false positive regression is run on the subset of all recidivists, and the false negative model is fitted on the subset of all non-recidivists.

Indeed, the unfavourable discrepancies in error rates for African-American defendants persist when adjusting for other relevant variables. Namely, the regression output suggests that African-American defendants are $e^{0.5431} \approx 1.72$ times more likely to be misclassified as high-risk criminals and $e^{-0.4137} \approx 0.66$ times as likely to be incorrectly classified as low-risk than Caucasian defendants, c.p.. This, combined with the results presented in Section 5.1.4 make a stronger case as evidence of COMPAS' unfair bias against African-American defendants.

Furthermore, based on the current regression results no conclusions can be drawn with respect to the statistical relationship between error rates and gender, as the coefficient estimates of sex_female are insignificant in both cases.

	coef	**std err**	**z**	***p*-value**	**[0.025**	**0.975]**
const	-1.6686	0.102	-16.347	0.000	-1.869	-1.469
sex_female	0.1867	0.103	1.812	0.070	-0.015	0.389
age_cat_greater_than_45	-1.4005	0.136	-10.310	0.000	-1.667	-1.134
age_cat_less_than_25	1.3885	0.105	13.198	0.000	1.182	1.595
race_african_american	0.5431	0.096	5.669	0.000	0.355	0.731
race_other2	-0.4806	0.145	-3.309	0.001	-0.765	-0.196
priors_count	0.2884	0.017	17.150	0.000	0.255	0.321
c_charge_degree_m	-0.1601	0.091	-1.766	0.077	-0.338	0.018

Table 9: Logit regression coefficient estimates and corresponding statistics from regression of false negative errors on a set of covariates, for general recidivism. Model is fitted on the subset of recidivists among all individuals.

	coef	**std err**	**z**	***p*-value**	**[0.025**	**0.975]**
const	0.6857	0.108	6.346	0.000	0.474	0.898
sex_female	-0.2483	0.127	-1.961	0.050	-0.496	-0.000
age_cat_greater_than_45	1.3046	0.146	8.940	0.000	1.019	1.591
age_cat_less_than_25	-1.2275	0.110	-11.189	0.000	-1.442	-1.012
race_african_american	-0.4137	0.102	-4.073	0.000	-0.613	-0.215
race_other2	0.6138	0.152	4.046	0.000	0.316	0.911
priors_count	-0.2536	0.015	-17.234	0.000	-0.282	-0.225
c_charge_degree_m	0.4759	0.097	4.883	0.000	0.285	0.667

Table 10: Logit regression coefficient estimates and corresponding statistics from regression of false positive errors on a set of covariates, for general recidivism. Model is fitted on the subset of non-recidivists among all individuals.

5.2 Results of fairness analyses and bias mitigation

This subsection reviews the most important results obtained from algorithmic bias analyses using

4.6). The goal of this section is to demonstrate the applicability of the considered bias mitigation algorithms on a general probabilistic classifier. First, the COMPAS risk scores are assessed using various measures designed specifically for evaluating disparities between sub-populations. Then, a similarly unfairly biased probabilistic classifier is trained and corrected by several post-processing techniques. Finally, a visualisation of fairness-accuracy tradeoffs is displayed and discussed.

5.2.1 Experimental setup

The experimental setup of applying fairness metrics and bias mitigation algorithms is similar to that of a general machine learning research pipeline. The full dataset $\mathbf{X}$ (and corresponding predicted and true outcome labels $\hat{Y}$ and Y) suspected of being unfairly biased is partitioned into non-overlapping subsets for training, validation and testing, with respective proportions of 70%, 15% and 15%. A (possibly unfairly biased) classifier, concerned solely with maximising prediction accuracy, is trained on the training data. Using the classification model fitted on the training set, predictions are made on the validation and test sets and their accuracy and fairness levels are recorded. These values represent efficacy and equitability properties of the decision-making process *before* post-processing or bias mitigation is applied. The validation set is then used to fit the bias mitigation model. For (calibrated) equalised odds post-processing ((C)EOPP), this entails solving a linear program to find a new distribution of prediction scores to satisfy (relaxed) equalised odds and / or calibration, whereas reject option based classification (RObC) finds an optimal new classification threshold τ and corresponding margin width θ to simultaneously maximise prediction accuracy and approximately satisfy some predetermined fairness constraint. Finally, generalised performance and fairness are evaluated on the test set. The post-processed or transformed evaluation metrics of the validation set are also reported.

A classifier's performance is assessed using balanced accuracy (abbreviated in tables as $BACC$), because this measure takes into account both outcome classes (and error types) by averaging the true positive and true negative rates. Group fairness is assessed using Statistical Parity Difference (SPD), Disparate Impact Ratio (DIR), Average Odds Difference (AOD) and Equal Opportunity Difference (EOD). The formal definitions of these fairness measures are outlined in Section 4.5. In this recidivism context, SPD measures the difference in likelihood of being assigned the *favourable* prediction label (i.e., being labelled as low-risk), conditional on group membership. So a negative SPD with respect to race suggests that African-Americans are *less* likely than Caucasians to receive low-risk classifications. DIR uses the exact same information as SPD, but take the ratio of the group-conditional rates of being labelled as low-risk, where a value of smaller than one implies the privileged group is *more* likely to receive lower risk scores. The benefit of considering DIR is its interpretation as the relative odds of receiving the favourable prediction label. Note that both SPD and DIR fail to take error rates into account, and thus cannot serve as comprehensive fairness measures by themselves. This drawback is compensated by considering AOD, which measures the average difference in false *and* true positive rates between the privileged and unprivileged groups. A negative AOD suggests that the unprivileged group is disproportionately disadvantaged, as it implies that they are more likely to be incorrectly classified as high risk and / or less likely to be mislabelled as low risk. An AOD of zero suggests the classifier is fair w.r.t. the protected attribute in question. An obvious limitation of AOD is that it places two sometimes incompatible restraints on a decision-making model, which could lead to unacceptably poor performance of overly penalised classifiers. Its relaxed counterpart, EOD, is therefore also reported, as it only measures the group-specific difference in correctly being labelled as low-risk by the RPI, making it a more easily attainable fairness criterion.

Notice that these group-based fairness metrics are all related in varying degrees, in the sense that

solving for one typically leads to the others also converging towards their points of fair outcome. A special case of the generalised entropy indices, the Theil index (i.e., $\mathcal{E}(\alpha)\,|_{\alpha=1}$), is also reported in an effort to consult an unrelated measure of inequality. Subgroup decomposability of the Theil index also provides a measure of within-group inequality, denoted by $\mathcal{E}_W(1)$.

It is worth repeating that when discussing group-membership with respect to race or gender, African-Americans and females are referred to as the *unprivileged* groups for race and gender, respectively. Whereas Caucasians and males are considered the *privileged* respective ethnic group and sex.

5.2.2 Evaluation of classifiers before bias mitigation

The COMPAS risk scores have certain limitations with regard to the applicability of the post-processing methods considered in this study. First, they are discrete values from 1 through 10. Second, the scores are not interpretable as (unnormalised) probability estimates of recidivism (i.e., a defendant with a COMPAS score of 6 is not considered to be "twice as likely" to recidivate as one with a score of 3). These aspects are problematic in the sense that any obtained classification threshold on the COMPAS decile scale (e.g., one found by RObC) doesn't have a probabilistic interpretation. Furthermore, there is also a considerable loss of information as a result of all continuous threshold estimates being set equal to the nearest integer. However, the fairness and balanced accuracy results of the COMPAS risk score are still discussed and displayed in Table 11.

The reasons mentioned above motivate the use of a probabilistic classifier to study the applicability of bias mitigation algorithms on a potentially biased decision-making process. A natural choice is a logistic classifier (see Section 4.3), as this model specification is both comparably accurate and easily interpretable from a probabilistic point of view. Table 12 contains the fairness metrics and balanced accuracy of the pre-bias mitigation logistic classifier for all three data subsets.

Attribute	**BACC**	$\boldsymbol{F}_1$	**SPD**	**DIR**	**AOD**	**EOD**	$\mathcal{E}(1)$	$\mathcal{E}_W(1)$
Race	0.657	0.675	-0.245	0.634	-0.207	-0.203	0.241	0.239
Sex	0.657	0.675	0.0449	1.088	0.005	-0.007	0.241	0.241

Table 11: Balanced accuracy, F_1 score, group-fairness metrics, and Theil indices on Broward County data set with COMPAS classifier. Unprivileged groups: African-Americans and Females.

Dataset	**Attribute**	**BACC**	**SPD**	**DIR**	**AOD**	**EOD**	$\mathcal{E}(1)$	$\mathcal{E}_W(1)$
Train ($\tau = 0.5$)	Race	0.670	-0.298	0.596	-0.261	-0.218	0.212	0.209
	Sex	0.670	0.246	1.483	0.218	0.173	0.212	0.209
Valid. ($\tau = \tau^*$)	Race	0.647	-0.188	0.743	-0.153	-0.098	0.188	0.188
	Sex	0.647	0.199	1.344	0.167	0.093	0.188	0.188
Test ($\tau = \tau^*$)	Race	0.651	-0.228	0.699	-0.201	-0.183	0.195	0.194
	Sex	0.651	0.209	1.361	0.164	0.149	0.195	0.195

Table 12: Accuracy (BACC) and fairness results of the logistic classifier (before post-processing): Privileged groups: Caucasians & Females. Optimal threshold: $\tau^* = 0.554$.

A quick comparison between Tables 11 and 12 shows both classifiers perform similarly with respect to balanced accuracy, all having values around 65%. The same can be said for both the direction and magnitude of the group-based fairness measures. Even in terms of the Theil index, the COMPAS

scores and logistic classifier display similar results. More explicitly, both classifiers appear to be severely disadvantageous towards African-American defendants. The COMPAS scores seem to be approximately unbiased towards any sex according to the group-based fairness metrics. Conversely, the logistic classifier seems to be very biased against male defendants, with positive values for SPD, AOD and EOD and $DIR > 1$.

5.2.3 (Calibrated) Equalised odds post-processing

Equalised odds post-processing (EOPP) yields inadequate results, as it consistently returns either the exact same classifier as the unprocessed one, or a trivial random guessing classifier that did satisfy the group-fairness optima. See Table 19 in the Appendix for an overview of these results. A possible explanation is that EOPP's requirement to satisfy the stringent equalised odds criterion by sheer prediction score reassignment is incompatible with the given probabilistic classifier and it therefore either returns the input or a trivial classifier that does in fact satisfy equalised odds, but is of course useless from a practical perspective. Incorrect model specification and deployment could also be a cause of these results. However, the most likely explanation is that the A-conditional ROC curves, across which EOPP searches for an optimal and feasible error rate tradeoff, only have trivial intersection points. As mentioned in Section 29, equalised odds post-processing is severely limited by its requirement of non-trivial intersections of A-conditional ROC curves. A lack of such intersection points would mean there is no feasible and reasonable error rate tradeoff to satisfy equalised odds. As a consequence, EOPP returns either the unprocessed (biased) classifier $\hat{Y}$ or the trivial classifier that does satisfy the fairness criteria, but is only as accurate as random guessing.

Calibrated equalised odds post-processing attempts to simultaneously satisfy calibration of a classifier and some predefined relaxation of equal error rates by finding a new classifier from the set of calibrated classifiers. This set of calibrated classifiers is equivalent to a line in the generalised FN-FP-plane connecting origin (i.e., $g_{fn}(r) = g_{fp}(r) = 0$ also known as the perfect classifier) and the calibrated trivial classifier (i.e., the classifier that returns the (group-specific) base rate for every input). This line is uniquely defined by the group-specific base rate. Satisfying calibrated equalised odds or a relaxation thereof corresponds to finding a set of sub-population-specific calibrated classifiers that lie on the same horizontal, vertical or level-curve defined by the decision-makers cost-weighing of false positive or negative errors. For a more detailed explanation of CEOPP, the reader is referred to Section 4.6.2 and the work by Pleiss et al. (2017).

CEOPP is applied to the logistic classifier with fairness and accuracy properties listed in Table 12. The model is fitted on the validation set for three types of costs constraints, with respect to race and sex separately. These cost constraints are approximately equal generalised false positive rates, generalised false negative rates or approximately lying on the same level-curve defined by equally weighted FPR and FNR. The fitted model is then tested on the test set. See Table 13 for an overview of these results.

Classification performance remains at slightly lower, but still acceptable levels compared to the pre-bias mitigation classifier. Furthermore, in terms of both balanced accuracy and fairness, using false negative rates or a weighted cost function of both error types yield superior results than only solving for false positive rates. This could suggest that ensuring that the rates at which defendants are falsely labelled low-risk are approximately equal across groups is a more stringent criterion on the classifier in question than attempting to equalise the rates at which defendants are mislabelled as high-risk.

Furthermore, classification performance on validation and test sets appears to be relatively similar, though fairness measures tend so stray further from their respective points of equal outcome when comparing validation set results to those of the test set. This pattern is line with the theory that

Dataset	Constraint	Attr.	$BACC$	SPD	DIR	AOD	EOD	$\mathcal{E}(1)$	$\mathcal{E}_W(1)$
Valid.	FPR	Race	0.5999	-0.5933	0.2089	-0.5628	-0.6763	0.3754	0.3160
		Sex	0.5208	-0.0246	1.6246	-0.0565	0.0561	0.1717	0.1564
	FNR	Race	0.6286	-0.4801	0.5102	-0.4653	-0.3221	0.1629	0.1518
		Sex	0.6531	0.4278	1.8099	0.4181	0.2827	0.1923	0.1864
	Weighted	Race	0.6352	-0.5310	0.4690	-0.5218	-0.3573	0.1646	0.1507
		Sex	0.6530	0.4755	1.9092	0.4759	0.3121	0.1879	0.1808
Test	FPR	Race	0.5883	-0.6368	0.1376	-0.6088	-0.7186	0.4015	0.3384
		Sex	0.5362	0.3856	3.3174	0.3523	0.4818	0.3730	0.3238
	FNR	Race	0.6400	-0.4830	0.4956	-0.4713	-0.3238	0.1803	0.1683
		Sex	0.6502	0.3462	1.6307	0.3371	0.2280	0.1919	0.1878
	Weighted	Race	0.6367	-0.5141	0.4859	-0.5062	-0.3375	0.1593	0.1468
		Sex	0.6582	0.4677	1.8845	0.4690	0.3004	0.1836	0.1768

Table 13: Accuracy and fairness metrics for calibrated equalised odds post-processing (CEOPP) under various optimisation constraints (false negative rate, false positive rate, and weighted FNR-FPR) on the validation and test data. Averaged over 100 draws.

classification accuracy and fairness tend to move in opposite directions, that is, accuracy tends to improve when fairness decreases and vice versa. It also suggests that overfitting could also occur with respect to fairness constraints, in the sense that fitting a fairness-constrained model to a specific dataset can generalise poorly when testing the same model on unseen data.

Most importantly, Table 13 clearly shows that CEOPP does little to improve for group-based fairness constraints, when comparing the fairness measures to the unprocessed classifier's results in Table 12. In fact, CEOPP seems to exacerbate disparities in almost all cases, depicted by group-based fairness metrics moving away from their equitable outcome points (e.g., the average EOD with respect to race for the test set goes from -0.183 to -0.3238 in the best case scenario). Moreover, CEOPP fails to satisfy the assigned cost constraint in most cases. For instance, when attempting to satisfy for approximately equal false negative rates (which is equivalent to having equal true positive rates), one would expect the equal opportunity difference metric to approach zero. This is clearly not the case when viewing the corresponding results. Even more so, CEOPP worsens the EOD in the FNR constrained case. One plausible explanation is related to Theorem 3, postulated by Kleinberg et al. (2016), suggesting that only perfect classifiers can satisfy calibration and (relaxed) equalised odds when base rates differ for disjoint sub-populations. Knowing that CEOPP selects classifiers from a set of calibrated classifiers, it is likely that the large significant differences in base rates between groups (see Figure 6) results in the infeasibility of finding a classifier that remotely satisfies a relaxation of equal error rates. This could lead the post-processing method to return comparatively accurate, calibrated classifiers with deteriorated error rate discrepancies, making them undesirable given the more equitable unprocessed original classifier.

It is worth mentioning that CEOPP does, however, reduce overall and within-group inequality, as measured by the Theil index. Although the improvements are subtle (the test-set $\mathcal{E}_W(1)$ drops by a factor of roughly 1.5 in the best case scenario), the opposite can be said for reject option based classification. As will be discussed promptly, RObC severely outperforms CEOPP in terms of between-group fairness, whilst exacerbating within-group disparities. This observation serves as a subtle, yet important illustration of the trade-offs between group- and individual-based fairness, discussed by Speicher et al. (2018), and Dwork et al. (2012), among others.

5.2.4 Reject option based classification

Reject option based classification (RObC) attempts to find a new optimal classification threshold, denoted τ^*, with margins of optimal width θ^*, with the goal of perserving classification performance whilst optimising a given fairness constraint. The method builds on the idea that individuals who lie near the decision boundary are most susceptible to unfair treatment, as their posterior class-membership probabilities convey the least certainty. RObC thus finds an optimal combination of a new decision boundary and corresponding margin width, where individuals who reside within the margins are classified according to a so-called reject option decision rule. More specifically, unprivileged defendants within the boundary are assigned the favourable label (i.e., categorised as low-risk), while privileged group members are assigned the unfavourable label (that is, classified as high-risk). See Section 4.6.3 or the work by Kamiran et al. (2012) for a more explicit overview of RObC's mechanism.

In this experiment, RObC is fitted to the validation data (i.e., τ^* and θ^* are obtained based on this dataset) by performing a grid-search over combinations of 1000 thresholds $\tau \in [0.01, 0.99]$ and 100 margin widths. RObC is applied with respect to both protected attributes, race and sex, and concerned with the dual objective of maximising balanced accuracy, while optimising a user-defined fairness constraint. The three considered fairness objective metrics are SPD, AOD and EOD. Contrary to (C)EOPP, reject option based classification leaves the posterior probability scores of the individuals unaltered and converges to a single new accuracy maximising decision threshold and fairness optimising (minimal) reject option margin, making it unaffected by the classification threshold used by the original unprocessed classifier.

The results of these experiments are presented in Tables 14 and 15, containing the balanced accuracy and corresponding group-based and individual-level fairness metrics, and optimal decision boundary and margin widths, respectively.

Dataset	Obj	Attribute	$BACC$	SPD	DIR	AOD	EOD	$\mathcal{E}(1)$	$\mathcal{E}_W(1)$
Valid.	SPD	race	0.665	-0.047	0.900	-0.011	0.000	0.259	0.258
		sex	0.529	0.049	2.229	0.038	0.041	0.611	0.604
	AOD	race	0.664	-0.042	0.909	-0.009	0.050	0.296	0.295
		sex	0.533	0.046	1.883	0.035	0.059	0.686	0.686
	EOD	race	0.643	-0.082	0.858	-0.045	0.023	0.225	0.224
		sex	0.622	0.097	1.350	0.077	0.017	0.370	0.370
Test	SPD	race	0.651	0.058	1.136	0.114	0.120	0.261	0.254
		sex	0.538	0.044	1.796	0.028	0.063	0.598	0.596
	AOD	race	0.639	0.025	1.057	0.058	0.037	0.302	0.300
		sex	0.541	0.033	1.584	0.024	0.024	0.681	0.680
	EOD	race	0.656	-0.061	0.893	-0.006	0.004	0.249	0.247
		sex	0.638	0.232	1.900	0.166	0.196	0.410	0.410

Table 14: Accuracy and fairness metrics for reject option based classification (RObC) under various optimisation constraints (statistical parity difference (SPD), average odds difference (AOD), equal opportunity difference (EOD), on the validation and test data.

RObC clearly reaches the desired goal of fairness and reasonable preservation of accuracy in the race-based classifiers cases, thereby convincingly outperforming CEOPP. However, when attempting to satisfy the given fairness constraints with respect to men and women, RObC, like CEOPP, fails to

Dataset	Fairness objective	Attribute	τ^*	θ^*
Valid.	SPD	race	0.420	0.068
		sex	0.182	0.000
	AOD	race	0.425	0.077
		sex	0.184	0.000
	EOD	race	0.480	0.048
		sex	0.334	0.000

Table 15: Optimal classification threshold (τ^*) and margin width (θ^*) for RObC, under various optimisation fairness-constraints, calculated on the validation dataset.

odds difference w.r.t. sex does RObC yield acceptable balanced accuracies for both the validation and test sets, with values around 63%. These desirable balanced accuracies are juxtaposed with poor fairness metrics, especially for the test set. The EOD solving classifier achieves very appealing levels of fairness, with SPD, AOD and EOD reaching zero, though contrasted by a disparate impact ratio of about 1.4, when evaluated on the validation set. Please note that $AOD = 0$ implies satisfaction of the stringent equalised odds criterion. But when the same classifier is confronted with unseen test data, the classification performance remains desirable, even slightly increasing, while the group-based fairness measure all skew towards biased levels against males. This suggests more evidence of fairness-overfitting, i.e., a fairness satisfying post-processing method fitting to a data-specific solution which doesn't generalise well.

The poor fairness and accuracy results with respect to sex are also reflected by the optimal parameters found for the corresponding RObC configurations. Namely, RObC consistently converges towards very low decision thresholds for classification as high-risk, i.e., the classifiers are labelling defendants as dangerous criminals at alarmingly low risk estimates. This phenomenon could be attributable to the large difference in recidivism prevalence among men and women, combined with the goal of deminishing disparaties against females.

As mentioned earlier, RObC does, however, perform well when reducing racial disparities among defendants. Balanced accuracy is often improved, compared to the pre-bias mitigation performance reported in Table 12, with negligible changes in accuracy between validation and test sets. And despite RObC's superior preservation of efficacy, group-fairness measures nearly reach their optima in most cases, alongside minimal increases in within-group equitability. Namely, $\mathcal{E}_W(1)$ increases by a factor of about 1.5 in the worst test scenario, which is concerned with solving for the most demanding fairness criterion, AOD. Surprisingly, the best generalised performing configuration of RObC (i.e., minimising the absolute value of EOD for race) not only achieves the highest balanced accuracy of all test cases, boasts the EOD closest to zero, but also yields the best value for AOD (even better than the configuration of RObC that attempts to optimise for this very metric). Recall that EOD is a relaxation of AOD. Though counter-intuitive at first sight, this observation could be caused by the over-constrained AOD-solving reject option based classifier failing to consider some apparently important scope of the problem space.

In terms of optimal thresholds and margin widths for racial disparity-minimising RObC, similarities can be seen between all three fairness objectives. The three values of τ^* lie somewhere between probabilities of 0.420 and 0.480 of perceived recidivism risk. But also in this comparison, EOD-solving RObC prevails over its race-concerned opponents, by achieving the highest optimal threshold with the lowest optimal margin width. Minimising margin width is appealing, as RObC's working premise, how effective it may be, involves switching prediction labels around, which is difficult to justify ethically. Thus, from a moral perspective, it is therefore advantageous to minimise the amount

of defendants who are subject to the reject option classification rule. A further discussion of this topic is saved for Section 6. Furthermore, a high optimal threshold is desirable for the defendant's perspective, however one could argue that from a societal well-being point of view, stricter decision boundaries would be preferable. A thorough investigation about these implications is best conducted in collaboration with an expert on law and criminology.
A final caveat with regards to RObC's apparent superior performance comes as a consequence of the previously mentioned Theorem 3, that states that equalised odds and calibration are unattainable in non-trivial cases. Seeing as how AOD reaches a value of approximately zero for the best performing configurations of RObC, proper calibration of the risk estimates of these classifiers is no longer possible. Meaning that the produced probability estimates carry different semantic meaning for Caucasians than for African-Americans, which is undesirable and considerable as unfair treatment (see Section 5.1.2).

5.2.5 Observed tradeoffs

Figure 9 displays a few noteworthy visualisations of observed tradeoffs between within- and between-group inequality measured by Theil indices (Figures 9c and 9d), and between balanced accuracy ($BACC$) and equal opportunity difference (EOD) for classifiers transformed by calibrated equalised odds post-processing with respect to approximate equality of weighted generalised error rates (9a and 9b). The left half of the plot-group correspond to the post-processed classifiers with respect to race and the right half to those corresponding with sex-disparity minimising classifiers.

In this experimental setup, the effects of altering the classification threshold of a probabilistic classifier on the balanced accuracy and EOD of the pre- and post-bias mitigation classifier. A total of 500 of equally distanced values of $\tau \in [0.01, 0.99]$ are considered, at each of which the balanced accuracy and EOD are plotted before and after applying CEOPP. EOD is chosen as a representation of group-fairness, because it uses error rate information, unlike SPD and DIR, which only look at group-specific rates of assignment of prediction labels. Moreover, EOD is a special case of AOD and hence more easily achievable. Furthermore, unlike AOD, the sign of EOD has a single clear interpretation: negative values indicate a disproportionately large amount of false negative errors for the unprivileged group and vice versa. Whereas the sign of AOD could be caused by discrepancies between groups in either or both of the error types.

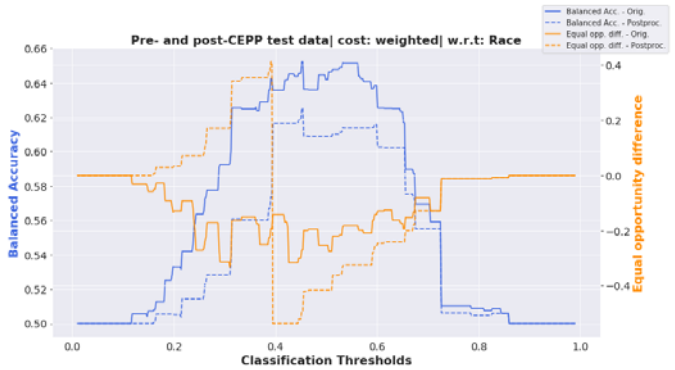
In the lower half of the plot-group in Figure 9, the total, within-, and between-group Theil indices are plotted for the same CEOPP configuration and set of classification thresholds as the plot directly above it. Recall that by sub-group decomposability of generalised entropy indices, $\mathcal{E}_T(1) = \mathcal{E}_W(1) + \mathcal{E}_B(1)$.

An immediately clear and reoccurring observation is the on average diminished balanced accuracy after applying CEOPP, depicted by the blue dashed curves in Figures 9a and 9b being lower than their solid blue counterparts at every threshold, showing that post-processing clearly constrains predictive accuracy. Another clear pattern from these upper two plots are the upward and downward movements of balanced accuracy accompanied by increasing and decreasing *absolute* values of EOD, illustrating the so-called fairness-accuracy tradeoff. This pattern persists before and after bias mitigation. A striking observation is the increased magnitude of unfairness, as well as a clear flip in direction of the bias after a certain threshold. The solid orange lines in Figures 9a and 9b display moderate unilateral deviation from the point equal opportunity (i.e., $EOD = 0$), in disfavour of African-Americans and males, respectively. Whereas the post-processed EOD values display extremely pronounced peaks in the opposite direction of the unprocessed bias for low-thresholds, before abruptly flipping over towards equally spiked levels of unfairness in the other direction at their respective bias-direction-thresholds. This failure to stabilise for equal TPR (except for trivial classifiers at the outer

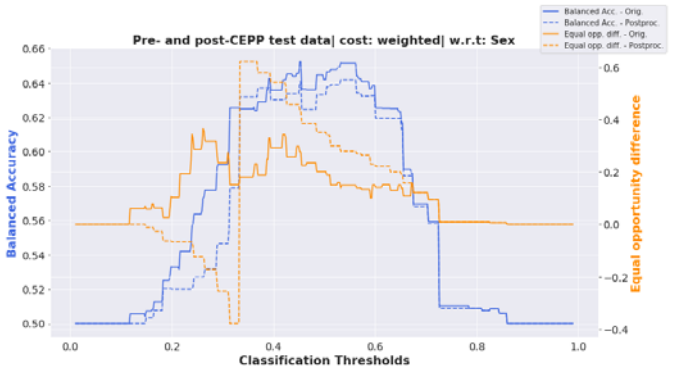
ends of the threshold range that amount to random guessing) accurately portrays the inability of calibrated classifiers to satisfy even the most lenient error-rate constraints. This also serves as a visual clarification for the poor results achieved by CEOPP, shown in Table 13, as it is apparently impossible for CEOPP to find a reasonable combination of efficacy and group-based fairness.

Finally, the lower two plots, Figures 9c and 9d, do not convey much information with regards to the tradeoffs in within-, and between-group inequality. What can be said, however, is that according to the Theil index, the vast majority of observed inequality is attributable to within-group disparaties. This observation persists after post-processing, but with a slight increase in between-group inequality. CEOPP only succeeds in lowering total inequality after the protected attribute-specific fairness-thresholds. If a decision-maker would be concerned with optimising for $\mathcal{E}_T(1)$ and $BACC$, CEOPP would produce reasonable classifiers for the higher part of the threshold range, but of course these models would result in severely undesirable error-rates disparaties.

It must be noted that the results presented in this subsection and specifically Figure 9 must be accepted tentatively, as CEOPP failed to produce either reasonable classification performance of fairness levels. The main purpose of these plots is to display empirical realisation of tradeoff patterns that are commonly cited in literature (Chouldechova, 2017; Corbett-Davies & Goel, 2018; Kamiran et al., 2012; Kleinberg et al., 2016). Furthermore, recall that RObC is unaffected by the threshold of a classifier, as it will converge to the same optimal threshold and margin for every starting threshold. Meaning that a fairness-accuracy plot of RObC would be a horizontal line at the optimal threshold. For that reason, I have chosen to omit such a graph.



(a) Line plots of balanced accuracy and equal opportunity difference before and after applying calibrated equalised odds post-processing w.r.t. race.



(b) Line plots of balanced accuracy and equal opportunity difference before and after applying calibrated equalised odds post-processing w.r.t. sex.

(c) Line plots of decomposed Theil indices before and after applying calibrated equalised odds post-processing w.r.t. race.

(d) Line plots of decomposed Theil indices before and after applying calibrated equalised odds post-processing w.r.t. sex.

Figure 9: Line plots of tradeoffs between balanced accuracy and equal opportunity difference, and decomposed inequality measured by Theil indices, before and after calibrated equalised odds post-processing for weighted error rates.

6 Points of discussion & topics for further research

Any informed debate about the implementation of lawful fairness in machine learning-guided decision-making systems can be the source of a plethora of topics of discussion about the moral implications of fair machine learning (ML). The most salient points of discussion and ideas for future research projects arising from the findings of this thesis are briefly touched upon in this section. Furthermore, this section covers some important considerations regarding the validity of the results reported in this paper.

One of the most fundamental moral dilemmas in fair ML is whether decision makers (e.g., judges, employers or creditors) should incorporate observed differences in base rates in their modelling process, thereby treating individuals differently based in part on their group-membership. For instance, if recidivism prevalence is observed to be significantly more pronounced among men or African-Americans, should these groups then be held to different standards than their respective opposite groups? On the one hand, this would avoid the issue of women being held to the same standard as men regarding recidivism risk scores, despite women showing significantly lower rates of recidivism, thereby reducing the rate at which women are incorrectly labelled as high-risk and unjust sentencing. Conversely, doing so with respect to race would imply subjecting new African-American defendants to harsher scrutiny in the form of different, more strict risk assessments than Caucasians. A consequence of such policy is the exacerbation of base rates differences and social disparities by creating a type of vicious cycle that becomes all the more difficult to break. More specifically, treating high-risk groups with stricter estimates is likely to increase the rate at which the group is exposed to risk, due to, e.g., increased surveillance, incarceration rates (which have known societal impacts that can lead to rising crime rates), and greater chances of being falsely labelled as high-risk.

This leads to a related fundamental subject of debate. Namely, the bias introduced via a target variable that is inherently incomplete, e.g., *observed* recidivism. It is likely that due to increased police surveillance in low-income, high-crime, and commonly ethnically diverse neighbourhoods, the odds of observing recidivism for racial minority groups is larger than for privileged groups. This phenomenon sets off a feedback loop that increases recidivism rates not only by actual rising crime rates, but also by the increased probability of unlawful behaviour being detected by the mere presence of law enforcers. This pattern is translatable to many different applications, e.g., consistently denying loan-applications of a minority group reduces the number of opportunities to work towards financial prosperity and become more creditworthy.

Narrowing the discussion to the bias mitigation results discussed in Section 5, an important debate arises from the intuition behind the best performing post-processing method, reject option based classification (RObC). In short, RObC finds an optimal decision-boundary for the estimated posterior probability of, say, a defendant becoming a recidivist with optimal margin width, within which the reject option classification rule holds that dictates that unprivileged group members are assigned the favourable label (e.g., classified as low-risk) and privileged defendants are given the unfavourable label. This process is similar to that of affirmative action, which satisfies a certain diversity quota by reassigning prediction labels, based purely on group-membership. For example, in an employment setting this sounds reasonable, though one could easily argue against accepting an unworthy candidate for a job. But when concerned with such high-stakes decisions as criminal sentencing, it is difficult to justify such behaviour of (seemingly) arbitrarily switching risk estimates, which in this application has immediate consequences in severity of penalties. Paradoxically, this intuitively odd mechanism performs very well in terms of preservation of balanced accuracy and optimisation of group-based fairness measures.

The intuition behind calibrated equalised odds post-processing (CEOPP) spawns similar moral

dilemmas. Contrary to RObC, CEOPP is subject to randomness, which is why the resulting fairness metrics and balanced accuracies are averaged over many draws. In practice, this would mean that, for instance, the results generated by a bias-corrected RPI for a specific defendant are non-deterministic, and can thereby vary from execution to execution. This also seems infeasible for judicial applications, as a defendant's sentence can be based on pure luck.

A further topic of discussion regarding RObC is the presumed sacrifice of calibration that comes with satisfying error rate equality across groups, as dictated by Kleinberg's theorem. That is, as a classifier achieves arguably important notions of fairness by satisfying equalised odds, the group-specific probability estimates lose their semantic meaning, in the sense that a risk estimate of a certain level would correspond to a different perceived risk of recidivism for members of one group than for members of another. The validity of these assertions, and a proper discussion of which fairness aspect weighs more heavily is left as a topic for further research.

7 Conclusions

The aim of this thesis was to investigate to what extent the recidivism prediction instrument (RPI) known as COMPAS can be accused of being unfairly biased towards African-Americans and / or women. Furthermore, this study set out the study the effectiveness of various bias mitigation post-processing techniques, specifically (calibrated) equalised odds post-processing and reject option based classification, when attempting to debias a presumably unfair arbitrary probabilistic classifier, whilst preserving prediction performance to obtain a reasonable, non-trivial, and fair classifier.

The experimental setup of this research was divided into two parts: assessing the unfairness of COMPAS scores using traditional non-parametric and parametric econometric methods, and applying the aforementioned bias mitigation algorithms to an arbitrary, but demonstrably biased probabilistic classifier.

The assertions regarding unfair treatment of African-Americans by COMPAS scores are confirmed with convincing statistical significance by logistic regression modelling of both the true and predicted outcome labels, as well as the false positive and negative error rates. These parametric modelling were, however, less conclusive with respect to gender-bias. While the rate at which women were labelled as high-risk (irrespective of correctness) are significantly higher than for men, compared to significantly lower rates of observed recidivism, the rates at which females were subject to mislabelling were insignificant in both cases.

When examining whether COMPAS scores are well-calibrated with respect to sex and race, this study found them to be calibrated for ethnic groups, but uncalibrated with respect to gender, in disfavour of females.

Furthermore, a non-parametric analysis of COMPAS scores by relative feature importance scoring using random forests showed a clear drop in relative importance of race-related variables when switching from predicting risk score categories to predicting actual observed recidivism. However, the opposite held for sex-related variables. Namely, they appeared to be more important relative to the other features when concerned with predicting observed recidivism.

As for the bias mitigation results, calibrated equalised odds post processing displayed very undesirable performance in terms of equitable outcome, often exacerbating the levels of both racial and gender disparities. Although CEOPP displayed acceptable conservation of classification performance for race-based post-processing, it failed to result in a reasonable improvement of the already biased classifier when aiming to reduce group-based inequality. A possible explanation is the impossibility of calibrated classifiers obtaining error rate equality.

Reject option based classification showed strikingly superior performance, both in terms of preser-

vation of efficacy and near optimal group-based fairness properties. However, it is highly likely that these newly obtained reject option based classifiers fail to satisfy calibration, which could be seen as unacceptable for practical purposes. Furthermore, it is disputable whether the intuition of affirmative action on which RObC builds is morally justifiable in a real-world criminal sentencing application.

Appendices

A Glossary

- **Anti-classification** A class of fairness definitions in which a decision-making model / algorithm does not consider protected attributes, like race, gender or proxies thereof when deriving estimates.
- **Bias** A systematic error in reasoning or logic that occurs as the result of the automaticity with which the human mind processes information based on expectations and experience
- **Bias mitigation algorithm** A procedure for reducing unwanted bias in training data or models.
- **Calibration** A fairness definition that requires outcomes of decision-making systems are independent of protected attributes after controlling for estimated risk. Put differently, an estimated risk score, e.g., $s(X)$, must correspond to the same risk for all individuals with that score, irrespective of their protected attributes.
- **Counterfactual** In causal inference and treatment evaluation, a counterfactual is an unobserved outcome that would have occurred, had the opposite decision been made of what actually happened. In a lending example, this corresponds to the unobserved outcome that would've been observed for a client whose loan application was denied, had it been accepted.
- **Decision rule** A decision rule is any measurable function $d : \mathbb{R}^p \mapsto \{0, 1\}$, where we interpret $d(x)$ as the probability that (binary) action a_1 is taken for an individual with visible attributes x.
- **Disparate impact** In US law, disparate impact refers to practices in employment, housing and other areas that adversely impact a protected group, despite the decision makers applied rules and intent thereof being neutral and non-discriminatory.
- **Disparate treatment** In US law, disparate treatment refers to unlawful discriminatory practices by decision makers towards an individual because of a protected characteristic.
- **Equalised odds** We say that a predictor $\hat{Y}$ satisfied equalised odds with respect to the protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y.
- **Fairness metric** A quantification of unwanted bias in training data or models.
- **Favourable label** A label whose value corresponds to an outcome that provides an advantage to the recipient, e.g., receiving a loan, getting hired, not being arrested.
- **Gerrymandering** The act of manipulating the boundaries of a population's subgroups, with the goal of benefiting a particular group or following some hidden agenda.
- **Group fairness** The goal of groups defined by protected attributes receiving similar treatments or outcomes.
- **Individual fairness** The goal of similar individuals receiving similar treatments or outcomes.

- **Infra-marginality** A general phenomenon in economics and statistics known as the problem of infra-marginality, refers to differences in error metrics across groups of protected and unprotected individuals due to their respective risk distributions being different.

- **Oblivious** A property of a predictor $\hat{Y}$ or score R (in supervised learning) is said to be oblivious if it only depends on the joint distribution of $(Y, A, \hat{Y})$ or (Y, A, R), respectively. Here, A denotes the (binary) protected attribute, whereas Y denotes the (binary) target variable of interest. Note that this property implies a model is oblivious to the information carried in X, the matrix of observable features.

- **Parity** A class of formal fairness definitions that requires certain predictive measures (often derived from a confusion matrix, like precision, recall, F1-score, FPR) be equal across all groups of individuals.

- **Protected attributes** An individual?s characteristics or attributes on the basis of which discrimination may occur. Examples include race (including colour, national or ethnic origin, or immigrant status), sex (including pregnancy or marital status and breastfeeding), age, disability, or sexual orientation, gender identity and intersex status.

- **Proxy** Data used to approximate labels or features not directly available in a dataset or not allowed to be used due to legal restrictions.

- **Sensitive attributes** See "protected attributes."

- **Sensitivity** Also known as the true positive rate. That is, the proportion of positive classes correctly classified as such by a binary classifier.

- **Specificity** Also known as the true negative rate. That is, the proportion of negative classes correctly classified as such by a binary classifier.

- **Subgroup validity** The phenomenon that certain features (say, x) are biased in the sense that factors are not equally predictive across (e.g., race) groups.

B Exhaustive list of variables and their descriptions

#	Name	Description	Type
0	id	Defendant's identification number in dataset	Positive integer
1	name	Defendant's full name	String
2	first	Defendant's first name(s)	String
3	last	Defendant's surname	String
4	compas_screening_date	Date of defendant's COMPAS assessment	Date (yyyy-mm-dd)
5	sex	Defendant's sex (M/F)	Categorical ($k = 2$)
6	dob	Defendant's date of birth	Date (yyyy-mm-dd)
7	age	Defendant's age at time of screening	Non-negative integer
8	age_cat	Defendant's age category	Categorical ($k = 3$)
9	race	Defendant's ethnicity / race	Categorical ($k = 6$)
10	juv_fel_count	Number of juvenile felony charges	Non-negative integer
11	decile_score	COMPAS risk decile score	Categorical ($k = 10$)
12	juv_misd_count	Number of juvenile misdemeanour charges	Non-negative integer
13	juv_other_count	Number of other juvenile charges	Non-negative integer
14	priors_count	Number of prior charges	Non-negative integer
15	days_b_screening_arrest	Days between arrest and COMPAS assessment	Integer
16	c_jail_in	Start of incarceration	Date-time
17	c_jail_out	End of incarceration	Date-time
18	c_case_number	Unique registered COMPAS case number	Categorical ($k = N$)
19	c_offense_date	Date of offense	Date (yyyy-mm-dd)
20	c_arrest_date	Date of arrest	Date (yyyy-mm-dd)
21	c_days_from_compas	...	Non-negative integer
22	c_charge_degree	Degree of offense	Categorical ($k = 2$)
23	c_charge_desc	Description of charge	String
24	is_recid	Whether previously registered by COMPAS	Binary
25	r_case_number	Recid case number	Categorical
26	r_charge_degree	Recid charge degree	Categorical
27	r_days_from_arrest	Days between previous arrest & screening	Non-negative integer
28	r_offense_date	Date of previous offense	Date (yyyy-mm-dd)
29	r_charge_desc	Description of previous charge	String
30	r_jail_in	Start of previous incarceration	Date (yyyy-mm-dd)
31	r_jail_out	End of previous incarceration	Date (yyyy-mm-dd)
32	violent_recid	...	...
33	is_violent_recid	Whether previously charged for violent crime	Binary
34	vr_case_number	Variable 25 for violent offense	Categorical
35	vr_charge_degree	Variable 26 for violent offense	Categorical
36	vr_offense_date	Variable 28 for violent offense	Date (yyyy-mm-dd)
37	vr_charge_desc	Variable 29 for violent offense	String
38	type_of_assessment	Type of COMPAS assessment	Categorical
39	decile_score.1	...	...
40	score_text	Decile score category	Categorical ($k = 3$)
41	screening_date	...	...
42	v_type_of_assessment	Type of COMPAS assessment (violent)	Categorical

43	v_decile_score	Violent recidivism risk decile score	Categorical ($k = 10$)
44	v_score_text	Violent decile score category	Categorical ($k = 3$)
45	v_screening_date	Violent recid risk screening date	Date (yyyy-mm-dd)
46	in_custody	Start of pretrial custody	Date (yyyy-mm-dd)
47	out_custody	End of pretrial custody	Date (yyyy-mm-dd)
48	priors_count.1	...	...
49	start	...	...
50	end	...	...
51	event	...	...
52	two_year_recid	Whether recidivated within 2 years	Binary

Table 16: Exhaustive list of variables that comprise the two primary data sets considered for analyses. Conventions: k denotes number of categories of categorical variable, "recid" is used as abbreviation for "recidivist."

C Proofs

C.1 Proof of the information theoretic data processing inequality

See Theorem 1.

Proof. By the Chain Rule, the Markov Chain's mutual information, $I(X,(Y,Z))$, can be decomposed in the following two ways:

$$\begin{aligned} I(X,(Y,Z)) &= I(X,Z) + I(X,Y \mid Z) \\ &= I(X,Z) + I(X,Z \mid Y). \end{aligned}$$

By assumption, $X \perp Z \mid Y \Rightarrow I(X,Z \mid Y) = 0$, and thus we obtain $I(X,Z) + I(X,Y \mid Z) = I(X,Y)$. Because the mutual information of two random variables is always non-negative, the data processing inequality follows from the previous expression: $I(X,Z) \leq I(X,Y)$.

□

C.2 Proof of Chouldechova's Incompatibility Result

See Theorem 2.

Proof. First, it is shown that base rates (μ), positive predictive value (PPV), and false positive and negative rates (FPR and FNR, respectively) are related via a single equation following from the definition of PPV:

$$\begin{aligned} PPV &= \frac{tp}{tp+fp} = \frac{TPR \cdot \mu}{TPR \cdot \mu + FPR \cdot (1-\mu)} \\ &= \frac{(1-FNR) \cdot \mu}{(1-FNR) \cdot \mu + FPR \cdot (1-\mu)} \\ &\Leftrightarrow \frac{1}{PPV} = \frac{(1-FNR)\mu + FPR(1-\mu)}{(1-FNR)\mu} = 1 + \frac{FPR(1-\mu)}{(1-FNR)\mu} \\ &\Leftrightarrow FPR = \frac{\mu}{1-\mu}\frac{1-PPV}{PPV}(1-FNR) \end{aligned}$$

Using the final equation, the theorem can be proven by contradiction. Assume there exist two disjoint sub-populations defined by protected attribute $A \in \{a,b\}$ with unequal base rates, and some binary classifier $\hat{Y}$ that satisfies predictive parity: $PPV_a = PPV_b$ and $\mu_a \neq \mu_b$.

Now assume the following to be true:

$$(FNR_a = FNR_b) \wedge (FPR_a = FPR_b). \tag{42}$$

Then the following must hold:

$$
\begin{aligned}
FPR_a &= FPR_b \\
&\Leftrightarrow \frac{\mu_a}{1-\mu_a}\frac{1-PPV_a}{PPV_a}(1-FNR_a) = \frac{\mu_b}{1-\mu_b}\frac{1-PPV_b}{PPV_b}(1-FNR_b) \\
&\Leftrightarrow \frac{\mu_a}{1-\mu_a} = \frac{\mu_b}{1-\mu_b}
\end{aligned}
$$

However, the last equation yields a contradiction as $\mu_a \neq \mu_b$. It must therefore hold that $FNR_a \neq FNR_b$ and/or $FPR_a \neq FPR_b$.

□

D Logistic regression results for violent recidivism

	coef	std err	z	P> |z|	[0.025	0.975]
const	-1.6705	0.095	-17.550	0.000	-1.857	-1.484
sex_female	0.2063	0.097	2.116	0.034	0.015	0.397
age_cat_greater_than_45	-1.3907	0.126	-11.034	0.000	-1.638	-1.144
age_cat_less_than_25	1.4274	0.097	14.787	0.000	1.238	1.617
race_african_american	0.5517	0.088	6.267	0.000	0.379	0.724
race_other2	-0.4591	0.132	-3.470	0.001	-0.718	-0.200
priors_count	0.2943	0.015	19.295	0.000	0.264	0.324
c_charge_degree_m	-0.2277	0.083	-2.745	0.006	-0.390	-0.065
two_year_recid	0.8605	0.108	7.958	0.000	0.649	1.072

Table 17: Model specification and parameter summary of logistic regression with score category (low or not low) as dependent variable for violent recidivism. Note that the variable race_other2 corresponds to a grouped dummy of all non-white and non-black ethnic groups.

	coef	std err	z	P> |z|	[0.025	0.975]
const	-2.0844	0.108	-19.296	0.000	-2.296	-1.873
sex_female	-0.6373	0.133	-4.794	0.000	-0.898	-0.377
age_cat_greater_than_45	-1.0749	0.147	-7.336	0.000	-1.362	-0.788
age_cat_less_than_25	0.6555	0.109	5.989	0.000	0.441	0.870
race_african_american	0.2446	0.105	2.326	0.020	0.038	0.451
race_other2	-0.1730	0.153	-1.128	0.259	-0.474	0.128
priors_count	0.1481	0.011	14.035	0.000	0.127	0.169
c_charge_degree_m	0.0539	0.096	0.562	0.574	-0.134	0.242

Table 18: Model specification and parameter summary of logistic regression with observed recidivism within two years as dependent variable for violent recidivism. Note that the variable race_other2 corresponds to a grouped dummy of all non-white and non-black ethnic groups.

E Equalised odds post-processing results

Dataset	Attribute	BACC	SPD	DIR	AOD	EOD	$\mathcal{E}(1)$	$\mathcal{E}_W(1)$
Valid. ($\tau = \tau^*$)	Race	0.647	-0.188	0.743	-0.153	-0.098	0.188	0.188
	Sex	0.647	0.199	1.344	0.167	0.093	0.188	0.188
Test ($\tau = \tau^*$)	Race	0.651	-0.228	0.699	-0.201	-0.183	0.195	0.194
	Sex	0.651	0.209	1.361	0.164	0.149	0.195	0.195

Table 19: Equalised odds post-processing

References

Altman, A. (2016). Discrimination. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2016/entries/discrimination/`.

Angwin, J., Kirchner, L., Mattu, S., & Larson, J. (2016). *Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks., may 2016.* ProPublica.

Bain, L. J., & Engelhardt, M. (1987). *Introduction to probability and mathematical statistics*. Brooks/Cole.

Barenstein, M. (2019). Propublica's compas data revisited. *arXiv preprint arXiv:1906.04711*.

Barocas, S., & Hardt, M. (2017). Translation nips tutorial: Fairness in machine learning. In *Proceedings of conference on neural information processing systems, long beach, usa.*

Barocas, S., Hardt, M., & Narayanan, A. (2018). *Fairness and machine learning*. fairmlbook.org. (`http://www.fairmlbook.org`)

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... others (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, *13*(2), 193–216.

Blais, J. (2015). Preventative detention decisions: Reliance on expert assessments and evidence of partisan allegiance within the canadian context. *Behavioral Sciences & the Law*, *33*(1), 74–91.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

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

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Crawford, K. (2016). Artificial intelligence's white guy problem. *The New York Times*, *25*.

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, *5*(2), 120–134.

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. *San Fransico, CA: Reuters. Retrieved on October*, *9*, 2018.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc.*

Dua, D., & Graff, C. (2017). *UCI machine learning repository.* Retrieved from `http://archive.ics.uci.edu/ml`

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226).

Everett, R. S., & Wojtkiewicz, R. A. (2002). Difference, disparity, and race/ethnic bias in federal sentencing. *Journal of Quantitative Criminology*, *18*(2), 189–211.

Fabian, J. (2012). The adam walsh child protection and safety act: Legal and psychological aspects of the new civil commitment law for federal sex offenders. *Clev. St. L. Rev.*, *60*, 307.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 259–268).

Feller, A., Pierson, E., Corbett-Davies, S., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. itâs actually not that clear. *The Washington Post*.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York.

Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *Nips symposium on machine learning and the law* (Vol. 1, p. 2).

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315–3323).

Hu, L., & Chen, Y. (2018). A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 world wide web conference on world wide web* (pp. 1389–1398).

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale JL & Tech.*, *18*, 148.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. In *2012 ieee 12th international conference on data mining* (pp. 924–929).

Kim, M., Reingold, O., & Rothblum, G. (2018). Fairness through computationally-bounded awareness. In *Advances in neural information processing systems* (pp. 4842–4852).

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237–293.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in neural information processing systems* (pp. 4066–4076).

Lieber, R. (2009). American express kept a (very) watchful eye on charges. *New York Times*, *30*.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*.

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Monahan, J. (1982). Clinical prediction of violent behavior. *Psychiatric annals*, *12*(5), 509–513.

Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. In *Proc. conf. fairness accountability transp., new york, usa.*

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 560–568).

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in neural information processing systems* (pp. 5680–5689).

Singh, J. P., Desmarais, S. L., Hurducas, C., Arbach-Lucioni, K., Condemarin, C., Dean, K., ... others (2014). International perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *International Journal of Forensic Mental Health*, *13*(3), 193–206.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2239–2248). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/3219819.3220046` doi: 10.1145/3219819.3220046

Stevenson, M. (2018). Assessing risk assessment in action. *Minn. L. Rev.*, *103*, 303.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

van Marle, H. J. (2002). The dutch entrustment act (tbs): its principles and innovations. *International journal of forensic mental health*, *1*(1), 83–92.

Waddell, K. (2016). How algorithms can bring down minorities' credit scores. *The Atlantic*, *2*.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web* (pp. 1171–1180).

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.