A Welfare Economics Approach to Combating Algorithmic Bias
an empirical analysis using fairness intervention techniques

Janssen, P.J.R.

Award date:
2021

A Welfare Economics Approach to Combating Algorithmic Bias:

AN EMPIRICAL ANALYSIS USING FAIRNESS INTERVENTION TECHNIQUES

Supervisors

dr. B.M. Sadowski	Industrial Engineering & Innovation Sciences
dr. C. Zednik	Industrial Engineering & Innovation Sciences
dr. G. Papachristos	Industrial Engineering & Innovation Sciences
ir. J. Veldman	Dialogic Innovatie & Interactie

Author	Patrick J.R. Janssen
Student number	0849091
Student email	p.j.r.janssen@student.tue.nl
University	Eindhoven University of Technology
Department	Industrial Engineering & Innovation Sciences
Master program	Innovation Sciences
Date	Wednesday January 20th, 2021

Acknowledgements

This report represents my master thesis project for the master program Innovation Sciences at Eindhoven University of Technology, with help from Dialogic Innovatie & Interactie in Utrecht. It was written over a period of 17 months. I would like to use this opportunity to thank the people who have helped me throughout this long and arduous journey, listed here in no particular order.

First, I would like to thanks my primary supervisor, dr. Bert Sadowski, for his guidance and unwavering faith in my abilities. Whenever I was stuck in my code, he helped me to take a step back and focus on the broader context. I would also like to thank my other supervisors, Carlos Zednik and Georgios Papachristos, for helping me complete my thesis even though we involved them only at a late stage.

Next, I would like to show my gratitude for everyone at Dialogic, and especially Jasper Veldman. Even though my final research topic was not in line with what we first envisioned, I appreciate that you let me work on a topic that was within my own interests.

I also want to thank everyone in my team at Rabobank, for letting me be flexible in my working hours and priorities. Without this support, I would have not been able to combine a full-time job with finishing this thesis during these last months. A special shoutout to my (former) talent manager Sandra Luijken for pushing me to be my best self and finish my thesis.

Lastly, thanks to all my friends and family, for putting up with my absence from social activities for 1.5 years, for daily coffee breaks and for stand-up meetings to keep me motivated. And of course to my girlfriend, not just for managing my work-life-thesis balance, but also for being my sparring partner in the field of algorithmic fairness. Without our discussions on my research design, I would not have been able to comprehend this report, let alone write it.

Abstract

Machine learning methods are increasingly used in decision-making processing, especially in screening decisions, such as hiring a new worker or giving out a loan. The inner workings of these methods are often opaque, which can lead to socially undesirable outcomes, e.g. screening out the wrong people. If the use of an algorithm leads to some population groups being treated unfairly, this is called algorithmic bias. State-of-the-art software tools exist to combat this bias by implementing bias mitigation techniques, which try to enforce parity over one or a combination of multiple fairness metrics, such as statistical parity or equality of opportunity. This often leads to decreased predictive accuracy for at least one of the groups. However, whether these bias mitigations actually lead to more societally desirable outcomes has not been researched yet. In this report, a method for empirically comparing algorithms that have been treated with several fairness interventions is developed. It uses a social welfare function to measure the societal preference of the outcomes of an algorithm. Results on binary classifiers trained on synthetically generated biased data show that treating algorithms with bias mitigation techniques leads to a decrease in both social welfare and predictive accuracy in 43% of the cases tested. This means that using the tools that have been designed to combat bias often leads to worse societal outcomes. This research aims to embed bias mitigation techniques within the context of welfare economics, providing insights into the trade-off between predictive accuracy and social welfare, as well modifying an existing methodology for determining the social desirability of different bias mitigation techniques.

Contents

Section 1: Introduction

1.1 Background

Spam filters, credit card fraud detection, search engines, news trends, advertising, insurance and loan qualification and credit scoring. These are only a subset of the commercial applications of machine learning algorithms people encounter daily. Increasingly, these algorithms are also used by governments to make decisions on a spectrum of topics, ranging from pretrial- (Dressel & Farid, 2018) and immigration detention (Koulish, 2016), to child maltreatment screening (Vaithianathan et al., 2013), public health (Potash et al., 2015), and welfare screening (Eubanks, 2018). These are all examples of *screening decisions*, wherein an algorithm is used to select one or more people from a larger pool based on an (unobserved) outcome of interest, and they can have large personal impacts on the people involved. Moreover, by using machine learning in these applications, these decisions are vulnerable to algorithmic bias; systemically creating unfair outcomes for some group of individuals (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2019).

Machine learning models work by extracting patterns from training data (learning) and using these to make predictions. The component that sets them apart from traditional models, is that they are not explicitly programmed. This allows machine learning models to capture complexities human minds are unable to think of. Machine learning algorithms can be classified as either supervised learning, unsupervised learning, or reinforcement learning. In supervised algorithms, the target variable is known. Examples are predicting housing prices, the weather, or customer turnover. Unsupervised algorithms, however, are more suitable for tasks such as clustering or dimensionality reduction. The third type of algorithm is reinforcement learning, in which an agent must make decisions on which actions it should perform in a simulated environment with the goal of getting some kind of reward. These types of algorithms can further be divided into which problem they tackle. The most important of these are classification (mapping input to labels or categories) and regression (mapping input to continuous output), which are both forms of supervised learning.

There are many types of machine learning models, all of which have their own sets of advantages and disadvantages. Examples of models include decision trees, logistic regression, support vector machines and Bayesian networks. However, the most recent advances have come in artificial neural networks. This field is called deep learning. Neural networks have found applications in fields such as computer vision and natural language processing (Belinkov & Glass, 2018), but also recommender systems, financial fraud detection, drug discovery, genomics (Lecun, Bengio, & Hinton, 2015) and customer relationship management (Tkachenko, 2015). They are also capable of generating fake

videos (Nguyen, Nguyen, Nguyen, Nguyen, & Nahavandi, 2019), fake art (Gatys, Ecker, & Bethge, 2016) and portraits of fake people (Karras, Laine, & Aila, 2018). This boom in use of neural networks has largely been attributed to increased availability of large datasets and increased computing power, which in turn has enabled models to become bigger, and thus, more accurate (Goodfellow, Bengio, & Courville, 2016).

This accuracy, however, comes at the cost of transparency. Large, deep neural networks can often contain millions of trainable parameters. How they exactly function and why they give certain outputs is often a riddle even to their creators. This opaqueness can potentially obscure socially undesirable mechanisms within algorithms, which can lead to sub-optimal outcomes.

Consider the widely publicized case of COMPAS, a risk tool that was used in the United States by judges to predict recidivism in order to determine pre-trial detention using machine learning algorithms. Ideally, COMPAS would predict with a high accuracy which people were most likely to commit a crime. However, a group of researchers found that the algorithm was biased against black people, putting them in a higher risk category for the same offenses when compared to people with different skin tones (Lansing, 2012; Larson, Mattu, Kirchner, & Angwin, 2016).

In this case, the algorithm used was subjected to an *audit*, to find out how and why they make decisions. However, there are a countless number of algorithms used worldwide, which are not subjected to these audits. This case illustrates that there is a need to audit algorithms, as this is the only way of knowing whether they function in the way they were intended to (Guszcza, Rahwan, Bible, Cebrian, & Katyal, 2018). By performing audits on machine learning algorithms used in consequential decision-making processes, users can be assured that these algorithms are safe, non-biased and fair. The General Data Protection Regulation (GDPR) has laid the legal basis for these audits (European Parliament & Council of the European Union, 2016; Schermer, Hagenauw, & Falot, 2018). Moreover, assuming that skin colour does not affect one's predisposition to commit crimes, if COMPAS had not been biased towards individuals with darker skin tones, it would also have been able to better predict the probability of committing crime, resulting in the right people being held in pre-trial detention. This, in turn, would lead to a decreased overall level of crime. This shows that algorithmic audits do not only lead to fairer outcomes but can also lead to more accurate predictions and thus better model performance.

1.2 Algorithmic audits

How such an audit would look like, is not a clear-cut case. Rambachan, Kleinberg, Ludwig, & Mullainathan (2020) argue that optimal regulation regarding algorithmic screening decisions exists, in which social welfare is maximized, because predictive accuracy is maximized and algorithmic bias is minimized, by implementing algorithmic audits to detect this bias. In the case of this optimal regulation, decision-makers (who use algorithms to assist in making a screening decision) can use all available characteristics of the individuals under consideration (including sensitive attributes such as race and gender). They are then required to publicly disclose their training data, trained algorithm and decision rule (the mapping from algorithm output to decision). This disclosure is referred to as an algorithmic audit. In reality, many predictive algorithms are immensely complex, and the concept of 'having access to the data, the model and the decision rule' is not as simple as sharing the code and the training data. Many types of machine learning models use stochastic processes that involve some degree of randomness and it is therefore impossible to exactly replicate such a model. Furthermore, trained algorithms and training data can be viewed as trade secrets and firms are unwilling to share any of them, in order to retain their competitive advantage.

This leads to auditors having to infer decision rules from a finite number of decision samples. Even then, when these decision rules are reconstructed, some notion of bias must be defined in order to test against. This is where the field of algorithmic fairness comes into play, which seeks to mathematically define notions of when an algorithm can be considered fair, i.e. not biased. There are many different statistical measures on which algorithms can be tested for fairness; do we want equal acceptance rates between different groups, or do we want similar individuals being treated similarly? Or do we want some other classification metric, such as true negative rate or positive predictive value to be equal among groups? Often, these different notions of fairness are incompatible with each other and cannot be satisfied simultaneously.

There are currently no accepted industry standards or best-practices concerning algorithmic audits. There are a few examples of fairness audits for algorithms performed by researchers (Chen, Ma, Hannák, & Wilson, 2018; Chouldechova et al., 2018; Edelman, 2011; Eslami, Aleyasen, Karahalios, Hamilton, & Sandvig, 2015; Eslami, Rickman, et al., 2015; Hannák et al., 2017; Sapiezynski, Wilson, & Kassarnig, 2017; Vijayakumar, 2018), but they are isolated cases, and do not have a shared methodology. There is also a single company that audits algorithms for fairness, but they do not explicitly explain their methodology (Winick, 2018). Furthermore, Auditing Algorithms: Adding Accountability to Automated Authority is a group of events designed to produce a white paper that

will define and develop the emerging research community for algorithmic auditing. To date, this paper has not been produced.

Recently, several algorithmic audit tools have been developed that are open source, meaning everyone is able to use them. The tools are software packages that take as input the training data and the predicted outcomes of the predictive algorithm and test this against some fairness definition. These include *audit-AI* by the start-up PyMetrics, *AI Fairness 360* by IBM, the *What-If Tool* implemented in TensorBoard by Google, *FairSight*, *FairVis*, *Aequitas* and *Fliptest* (Ahn & Lin, 2019; Bellamy et al., 2019; Cabrera et al., 2019; PyMetrics, 2017; Saleiro et al., 2018; Weinberger, 2018). In general, these tools can detect a number of different biases, according to a varying number of fairness definitions. Some, such as AI Fairness 360, are also able to then correct these biases, using several different techniques.

In screening decisions, a concept that can help in this situation is the *social welfare function*, which defines society's preferences over the *outcomes* of the screening problem (Rambachan, Kleinberg, Ludwig, & Mullainathan, 2020). An audited algorithm that has been judged to be biased according to a fairness definition can then receive a *bias mitigation* or *fairness intervention* that essentially makes the algorithm fair. There are many techniques for doing this and it remains an active field of study (Mehrabi et al., 2019). Using a social welfare function to calculate social welfare scores for algorithms that have been subjected to such a fairness intervention allows for comparison of the societal preference for using certain fairness definitions to audit algorithms.

The problem that this research addresses is that bias mitigation techniques try to make algorithms fairer according to certain fairness definitions, thereby altering the decisions made using these algorithms. However, it is unclear whether this also leads to more societally desirable outcomes. The aim of this research is to address this gap by empirically comparing algorithms that have been treated with several different fairness interventions. How societally desirable the outcomes of these mitigated algorithms are, is captures by a social welfare function. This research contributes to existing research in several ways. First, an overview of the existing literature on algorithmic bias in the context of screening decisions, and how to combat this bias, is created. Secondly, different fairness interventions are theoretically embedded in the context of social welfare. Thirdly, a method, adapted from the works of Rambachan et al. (2020) and Zafar, Valera, Rodriguez, & Gummadi (2017), for empirically calculating social welfare levels for different fairness interventions on synthetic datasets is developed. To my knowledge, the second and third contributions are novel relative to existing research.

The main research question that this report will be built around is **To what extent do fairness interventions influence the trade-off between social welfare and predictive accuracy in the context of algorithmic decision-making?**

1.3 Research structure

The report will be structured as follows. Section 2 will be a literature review on the existing research in the fields of algorithmic bias and algorithmic fairness. The first part will focus on how bias arises and its historic roots, answering sub-question 1, *What constitutes algorithmic bias in the context of screening decisions?* The second part of Section 2 will focus on the discussion on how a fair algorithm is defined, how algorithms can be made fair and what role the audit plays in this process, providing an answer to sub-question 2: *How can bias be detected and combated in algorithms?* In Section 3, the methodology used in this research will be described, focusing on how the social welfare function and the fairness definitions used are defined. It provides an answer to sub-question 3: *How can social welfare functions be used to quantify society's preferences of the outcomes of algorithmic decisions?* Section 4 will provide an analysis of the results, answering sub-question 4: *Is there a trade-off between social welfare and prediction accuracy in the context of bias mitigation techniques?* Finally, Section 5 will include the implications of the analysis, both practical and policy-related, as well as a discussion on the limitations of this research. It will strive to answer the main research question.

Section 2: Literature

2.1 Bias in algorithms

The following section aims to find an answer to **SQ1:** *What constitutes algorithmic bias in the context of screening decisions?* In order to do so, the context around screening decisions will first be introduced.

2.1.1 The screening problem

Algorithms based on machine learning are becoming increasingly popular options to support decision-making processes. As algorithms become more and more complex, so does their ability to capture complex patterns contained in data. These patterns can be exploited to create accurate predictions of any number of factors; predicting whether a candidate is suitable for a job, whether an obligor will default on his loan, whether a picture is of a dog or a cat, whether a chest X-ray image show signs of COVID-19 or whether a Facebook post contains hateful speech (Maguolo & Nanni, 2020; Ng, 2018). These factors are called *outcomes of interest*. A growing supply of data on which to train these models combined with increased available computing power has driven the rise of algorithmic decision making (Han, Kamber, & Pei, 2011).

2.1.2 Algorithmic decision-making

Decision-makers, such as a hiring manager at a firm, a judge overseeing pre-trial detention, or a credit analyst at a bank, then use the prediction made by a model to help make a decision; if a candidate is predicted to have a high productivity, then the candidate is hired; if an individual is predicted to be likely to recidivate, the individual is held in pre-trial detention; if a client is predicted to have a low creditworthiness, then the client does not get a loan (Chouldechova, 2017). The assumption here is that the predictive models learn which characteristics of an individual are related with the outcome of interest.

2.1.3 Discrimination, bias and fairness

Thus, the algorithms can help decision-makers discriminate between desired individuals and undesired individuals. This form of discrimination is valuable; an algorithm that is better at distinguishing between productive and unproductive workers has more economic value to a company hiring new workers. However, when using modern machine learning techniques, whose inner workings are often opaque, can lead to a situation in which characteristics that are not causally related to the outcome of interest, are used to make a prediction, such as race or sex. Even when

such characteristics are not used, the outcomes of the algorithm can be such that some group of individuals is preferred over another. When the use of these predictive model in the decision-making process leads to differential treatment between different protected attribute groups, then there is *bias*. The terms *discrimination* and *bias* are used interchangeably throughout the literature in the field of algorithmic decision-making. In this report, I will use *bias* to refer to *disparate impact*: a situation in which the use of an algorithm *unintendedly* creates unfair outcomes for some group of individuals, in line with Chouldechova (2017) and D'Alessandro, O'Neil, & Lagatta (2017). Thus, the concept of bias is linked to that of *fairness*. Mehrabi et al. (2019) define fairness in the context of decision-making as the *absence of any prejudice or favouritism toward an individual or a group based on their inherent or acquired characteristics.* Thus, fairness is measured on the *outcome* of the algorithm; of the decision that the algorithms helps to make. Bias, on the other hand, can be present in the data, in the algorithm trained on the data, or the outcomes of the algorithm.

The literature often contrasts disparate impact to *disparate treatment*, which refers to *intended* discrimination on the basis of protected attributes (Alexander, 1992; Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2014). See, for instance, Barocas & Selbst (2016) for a legal perspective on the difference between disparate impact and disparate treatment in the context of algorithmic decision-making.

2.1.4 Protected attributes

In Europe, which attributes can be used for algorithmic decision-making, and which ones are protected has been defined in de GDPR. The GDPR distinguishes special categories of personal data; *sensitive data* and *personal data relating to criminal convictions and offences*. Sensitive data are personal data revealing racial or ethnic origin, sexual orientation, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for identifying natural persons and health data. Generally, these data cannot be used or processed, only in highly exceptional cases. Exceptions include protecting the vital interests of the person involved in case the person is physically or legally unable to consent; when the data has been made public by the person; or when courts act based on their legal capacity. Processing criminal personal data is only allowed when under governmental supervision (European Parliament & Council of the European Union, 2016; Schermer et al., 2018). In summary, personal data can only be used under strict criteria, and even then, protected attributes cannot be used. As such, algorithmic discrimination based on these attributes should not be possible. However, research has shown that other data can be used as

proxies for protected attributes (Alexander, 1992). Therefore, excluding protected attributes from the data used does not guarantee fairness.

2.1.5 Sources of bias

So far, we have assumed that using machine learning algorithms in decision-making processes can lead to bias. Here, we will dive deeper into how this mechanism exactly works. Often, the term *biased data* is used, but in reality, disparate impact can arise at multiple points in the machine learning pipeline; in the data generation phase, in the model building phase and in the implementation phase. Mehrabi et al. (2019) and Olteanu, Castillo, & Diaz (2016) list a wide range of sources of bias. Here, I will follow Suresh & Guttag (2019) to indicate which (general) types of biases there are and how they can influence the decisions made by a model.

- *Historical bias*: Even if data is measured and sampled to perfectly represent the world as it is, including every relevant feature, a model built using this data can still lead to disparate impact. The world around us reflects historical issues such as prejudice and stereotyping. An example is given in Lum & Isaac (2016): there is empirical evidence that police officers, either consciously or not, consider race and ethnicity when deciding in which neighbourhoods to patrol and which people to apprehend (Lange, Johnson, & Voas, 2005). When certain neighbourhoods and ethnicities are more heavily policed, then these groups can be reasonably expected to be over-represented in police databases. This will result in a vicious circle, in which those ethnicities and neighbourhoods are more heavily policed because their crime rates are higher, because they are more heavily policed, and thus their *reported crime rate* will be higher.
- *Representation bias*: Representation bias can ensue when there is *selection bias*: the sample distribution (the data the model is trained on) does not match the true population distribution (the real world). This can happen when the sampling method is not inclusive of all groups, or when the population of interest does not match the training data. This type of bias can also transpire even if there is no selection bias; if some group is a minority, then the model will be less robust for this minority then for the majority group because there are fewer data points for the minority, resulting in less accurate predictions for this minority group. Less accurate predictions can, for instance, increase the chance for individuals to be wrongfully rejected.
- *Measurement bias*: Measurement bias arises when the features and labels that are used in a model are *proxies* for the actual features and labels under study. These proxies often have

different distributions for different groups, leading to different biases across groups. Suresh & Guttag (2019) identify three ways in which this can happen. Firstly, the *granularity* of data can vary across groups. Related to the policing example above, this happens when certain groups are monitored more often, leading to more observed errors for those groups, which can result in a feedback loop, as these groups are then monitored even more because of their higher error rate. To continue the above policing example, a model trained on this data does not actually predict crime, but some interaction between crime, policing strategy and taste-based discrimination (Lum & Isaac, 2016). Barocas & Selbst (2016) call this phenomenon *overrepresentation in the dataset*. Measurement bias can also occur because the *quality* of data differs across groups. This can lead to different results for members of different groups. The third way measurement bias arises is because the classification task itself is an oversimplification of problem at hand. Using machine learning methods to predict a certain label necessarily reduces the outcome to a single value. An example can be found in an algorithm used to screen job candidates. Whether a candidate should be hired for a job or not depends on many different factors, many which are not easily measured or expressed as a single value, such as fit within the company (culture), growth potential or leadership qualities. An algorithm used in this context is then expected to condense the suitability of a candidate into a single value, such as expected productiveness (expressed in a monetary unit). Other examples are using grades as a measurement for academic success, or rearrest rates as a proxy for crime.

- *Aggregation bias*: This form of bias occurs when one model is used to model multiple groups that have different conditional distributions. The mapping from inputs to outputs is often not the same for different groups. Forcing a single mapping for all groups can lead to a situation where a model does not adequately model the distribution of any subgroup. It is also related to representation bias, in the case that the model is fitted to the dominant group, leading to bias against minority groups.
- *Evaluation bias*: Evaluation bias is based on the need to objectively compare different models to each other. To do so, standardized benchmarks are often used. This leads to models being optimized for both their training data and the external benchmarks used and can also lead to overfitting on the benchmarks. This can be problematic if the benchmark is not representative of the problem at hand. This type of bias can also be aggravated by using a single metric (such as accuracy) to determine how well a model performs over all subsets, which can hide disparities in other types of errors.

2.2 Fairness in algorithms

Now that we have defined what constitutes bias in algorithms and how it arises, it is important to look to the field of *algorithmic fairness*. In the computer science literature, algorithms are often viewed not from a perspective of how they are biased, but whether or not they create *fair* outcomes for the individuals concerned.

2.2.1 Different fairness notions

Despite the prevalence of the term in the literature, there is no universally accepted definition of a fair algorithm. What constitutes fairness is a matter of debate, with a relatively large portion of the work in fairness has gone into mathematically formalizing various notions of fairness. Broadly, these notions of fairness can be split into the following categories, synthesized from overviews by Gajane & Pechenizkiy (2017), Naudts (2018) and Verma & Rubin (2018).

2.2.1.1 Group Fairness

Group fairness aims to treat different groups of individuals equally (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012; Kusner, Loftus, Russell, & Silva, 2017). Different notions exist on the exact metric that the groups should be compared on, including the following:

- *Group-independent predictions*: Under this notion, decisions made by an algorithm are fair when they are (conditionally) independent of sensitive group membership. It is also called *fairness through unawareness*, as the algorithm is unaware of group membership. However, this approach cannot guarantee discrimination-free practices, as features highly correlated with protected attributes (*proxies*) can still be included in the model (Pedreschi, Ruggieri, & Turini, 2008).
- *Statistical parity*: Also known as *demographic parity*, this notion requires an algorithm to predict an outcome for individuals across protected groups with almost equal probability. What sets this definition apart is its independence from the ground truth; the true distributions of the labels for different groups do not need to be known. This is useful in settings where this ground truth is not readily available, such as in settings where discrimination against protected groups is abundant, e.g. housing, employment, credit and consumer markets (Pager & Shepherd, 2008). It also enjoys legal support in the US, in the form of the *four-fifth* rule, which prescribes that the selection rate for a protected group cannot be lower than 80% of the highest rated group, in the context of employment (Greenberg, 1979). There is also evidence that introducing such criteria can have significant

benefits in building the reputation of the disadvantaged group in the long run (Hu & Chen, 2018). However, this approach is not suitable when the distributions for different protected groups do differ, i.e. when there is a correlation between protected attributes and the outcome (Luong, Ruggieri, & Turini, 2011). In this case, maintaining statistical parity may lead to discrimination against qualified candidates (Gajane & Pechenizkiy, 2017).

- *Equality of opportunity*: *Equality of opportunity*, which is the mostly used variant of *equalized odds*, is satisfied when the true positive rates for all groups are the same (Hardt, Price, & Srebro, 2016). Equivalently, *disparate mistreatment* stipulates that misclassification rates across groups are equal (Zafar et al., 2017). The downside of this notion is that it does not consider the effect of discrimination based on protected attributes. This discrimination can have an adverse effect on the opportunities one has access to, as is well documented in domains including education, business and politics (Gajane & Pechenizkiy, 2017). If the underlying distributions of labels are the same for all groups, then satisfying equality of opportunity also implies satisfying group fairness. However, when this is not the case, a classifier satisfying equality of opportunity will have a higher probability of assigning a positive value to the privileged group, thus dissatisfying the condition of group fairness.
- *Predictive rate parity*: The constraint of predictive rate parity is satisfied when both *positive predictive parity* and *negative predictive parity* are satisfied. This is the case when both protected and unprotected groups have an equal probability of a positive (or negative) prediction when they truly belong to the positive (or negative) class. For instance, in loan applications, this constraint is satisfied when both male and female, or white and non-white, applicants have the same probability of having a good (or bad) predicted credit score when they actually have a good (or bad) credit score (Verma & Rubin, 2018).

2.2.1.2 Individual Fairness

Individual fairness requires that individuals who exhibit similar characteristics are treated similarly, i.e. the algorithm predicts the same outcome. It essentially states that the distance between the outcomes of two individuals should not be greater than the distance between the features of the individuals (Dwork et al., 2012). However, the main critique of this approach is that such a distance metric, which measures the similarity between two inputs (individuals) is hard to determine (e.g. how do you quantify the distance between a bachelor's and a master's degree?)

A variant of individual fairness, *causal fairness* implies that a predicting algorithm is fair if the output remains the same when a protected attribute is flipped to its counterfactual value. This definition is also called *counterfactual fairness* (Kilbertus et al., 2017; Kusner et al., 2017; Kusner, Russell, Loftus,

& Silva, 2018; Loftus, Russell, Kusner, & Silva, 2018; Madras, Creager, Pitassi, & Zemel, 2019; Wu, Zhang, & Wu, 2019; Wu, Zhang, Wu, & Tong, 2019). Research has shown counterfactual reasoning to be susceptible to hindsight bias and outcome bias (Petrocelli & Sherman, 2010; Roese & Olson, 1996; Wu, Zhang, & Wu, 2019). This makes algorithms using this notion of fairness unsuitable for domains where these biases are frequently observed, such as judicial or healthcare systems (Gajane & Pechenizkiy, 2017).

2.2.1.3 Impossibility theorem

Research has shown the impossibility of satisfying the three most prevalent notions of group fairness simultaneously. These three are equalized odds, statistical parity and predictive rate parity. For an in-depth discussion and mathematical proof on this issue, see Chouldechova (2017), Kleinberg et al. (2017) and Miconi (2017). This means that a trade-off must be made between statistical parity, equality of opportunity and predictive rate parity. In practice, this means that no algorithm can ever be truly free of all types of bias, as fulfilling the requirements of one of these three definitions means violating the others. This is also the issue at the heart of the COMPAS debate referred to in the introduction. ProPublica argued that the COMPAS tool violates equality of opportunity, but Northpointe points out that the tool does satisfy predictive rate parity (Dieterich, Mendoza, & Tim Brennan, 2016; Lansing, 2012; Larson et al., 2016).

2.3 Solutions to the problem of bias in algorithms

For bias in algorithms to be combated, this bias first needs to be ascertained. The process of determining whether a predictive algorithm contains some sort of bias is called an *algorithmic audit*. The following section aims to provide an answer to **SQ2:** *How can bias be detected and combated in algorithms?*

2.3.1 Detecting bias: algorithmic audits

The General Data Protection Regulation (GDPR) has laid the legal basis for these audits. Under the (GDPR), personal data can only be processed if one if one the following criteria applies:

- The person involved consents to this use
- Processing is necessary for executing an agreement
- Processing is necessary for compliance with a legal obligation
- Processing is necessary to protect vital interests
- Processing is necessary for executing a task in the public interest or for a public authority

- Processing is necessary for the protection of legitimate interests

Furthermore, only data relevant to the task can be used. These criteria apply to all aspects of personal data. Some researchers also suggest the GDPR provides a *right to explanation*, which gives citizens the right to know how the algorithms that influence their lives work and how they make their predictions. How this right should be implemented however, and to what extent it is enforceable, is still under fierce debate; see, for instance Casey, Farhangi, & Vogl (2019) and Wachter, Mittelstadt, & Russell (2017). Opening these algorithms up in an audit could provide a solution to this debate.

The German Data Ethics Commission (Datenethikkommission) recently published a report on policy recommendations for algorithms and AI. In this report, they classify algorithms according to their risk potential, with 5 levels being identified. For level 1 algorithms, which have negligible potential for harm, no measures have to be taken, while level 5 algorithms (untenable potential for harm) should be banned altogether. As German laws were an influential input in the GDPR, these recommendations might someday be adopted into EU policy. In line with the recent work on fairness in algorithms, the recommendations call for transparency, explainability, accountability and fairness. With respect to auditability, they recommend that all regulated algorithms (risk levels 2 through 4) be exposed to audit procedures. Furthermore, they support all initiatives that develop quality standards for test procedures and audits, as well as advisory boards to ensure inclusivity of audits (Datenethikkommission, 2019). This supports the argument in Casey et al. (2019) that audits could become mandatory in the near-future.

2.3.2 Removing bias

When bias in an algorithm has been revealed by an audit, naturally the next step is removing it. Measures that enhance fairness by removing bias are known by many different names, but are most conventionally called *fairness(-enhancing) interventions* or *bias mitigation techniques*. These names will be used interchangeably in this report.

Researchers generally make a distinction between the methods used in removing bias based on where in the modelling pipeline they are used. In *preprocessing* methods, the input data to the algorithm is modified; using *inprocessing* methods, the algorithm itself is either designed to be fair from scratch, or an existing algorithm is modified to be fair; in *postprocessing* techniques, the output of the algorithm is modified so that the outcomes are more fair. Most fairness interventions target a single one of these categories (Friedler et al., 2019).

- *Preprocessing methods*: Preprocessing methods, alternatively known as data-based methods assume that the data underlying a machine learning model can be biased, and as the algorithm learns from the data, the algorithm thus also becomes biased. Modifying this data to be less biased will thus cause the algorithm to be fair. The cause of the bias in the data can the historical context in which it was generated (historical bias), forms of measurement error (measurement bias) or underrepresentation of certain groups (representation bias) (Friedler et al., 2019; Suresh & Guttag, 2019). A popular example of historical bias is present in recidivism models, i.e. algorithms designed to predict whether an inmate will commit another crime when they are released. These models are trained on data, however, that does not capture who commits crime but who is arrested for committed a crime. There is evidence that arrest rates are skewed towards minorities, as they face higher police rates (Rothwell, 2014). Thus, the algorithms learning from this biased data will predict higher recidivism rates for these minorities. One of the first methods for debiasing algorithms was also a preprocessing method (Feldman et al., 2014). A promising variant of preprocessing can be found in the work of Oneto, Donini, Maurer & Pontil (2019), who propose to use a fair representation of the data that can then be used in downstream modelling tasks. This representation does not include any sensitive attributes, decreasing the risk of leaking sensitive data (Shrestha & Yang, 2019).
- *Inprocessing methods*: The most common approach, algorithm modification, also called *algorithm modification* or model-based methods, place additional constraints on the learning algorithm, in order for it to not only optimize predictive accuracy but also a fairness criterium. Zehlike et al. (2017) created a fair algorithm for the top-K ranking problem.
- *Postprocessing methods*: Also known as post-hoc methods, these techniques modify the results of a trained predictive model so that the outcomes exhibit the desired fairness characteristics. These methods take as input the score output of a classifier and search for a threshold of each separate group so that some fairness metric is optimized. While this method does need access to the protected attribute, it has the advantage of being able to be applied to any trained classifier, as well as being computationally simple (Hardt et al., 2016; Shrestha & Yang, 2019).

The methods described above are all applicable to group fairness. Research into fair algorithms for individual fairness is scant and thus far uses too many assumptions to be useful in practice (Dwork et al., 2012; Joseph, Kearns, Morgenstern, Neel, & Roth, 2016; Shrestha & Yang, 2019). So far, research on fairness in the context of reinforcement learning has solely been focused on algorithm modification (Jabbari, Joseph, Kearns, Morgenstern, & Roth, 2017; Weng, 2019).

However, Kusner et al. (2018) rightly note that while recent work has been largely aimed at finding and removing biases, not much research has been performed in understanding which measures are right for a given problem. The proliferation of fairness interventions has not worked in this respect, as it is not even clear whether different measures actually differ from each other. Friedler et al. (2019) found that different fairness definitions that various fairness interventions try to combat are correlated, meaning that an algorithm designed to optimize a specific fairness measure can also be useful for other fairness measures. Causality-based methods can also provide an alternative, as they try to provide an understanding of the causal connection between protected attributes and decision, giving insight into *how* bias arises (Glymour & Herington, 2019; Kilbertus et al., 2017; Kusner et al., 2018; Loftus et al., 2018; Madras et al., 2019; Wu, Zhang, Wu, et al., 2019).

2.4 Optimal regulation of algorithms

Even though decision-makers, model builders and machine learning experts are becoming increasingly aware of the (unintended) adverse impacts their algorithms can have, the issues surrounding bias have not yet subsided. Therefore, it can be argued that algorithms need to be more strongly regulated, in order to force these users to create more fair algorithms (Doshi-Velez et al., 2017). The following section will give an answer to **SQ3:** *How can social welfare functions be used to quantify society's preferences of the outcomes of algorithmic decisions?* as it elaborates on what social welfare functions are and how they are used in economic settings.

2.4.1 Optimal regulation

Rambachan, Kleinberg, Ludwig, & Mullainathan (2020) describe a welfare-economics approach on the regulation of algorithms that will provide the framework for the analysis in this report. They define a theoretically optimal algorithmic regulation, in which bias regarding decision-making systems is reduced vis-à-vis a world in which all decisions are made solely by humans. This optimal regulation consists of decision-makers disclosing their predictive algorithm, the data the algorithm was trained on, and a *decision rule*, to a so-called *social planner*. The decision rule takes the prediction made by an algorithm and assigns it to a decision. Continuing the examples used before, it takes for instance a predicted credit score, and assigns a threshold above which clients' loans are approved, or it takes a predicted probability of recidivism, and defines a threshold above which defendants are detained. In the job application example, the decision rule is responsible for defining which predicted productivity level is sufficient for a suitable candidate. A social planner in this context is party that is concerned with optimizing social welfare and has control over public policy regarding algorithms. It can be viewed as a government institution responsible for regulating the

algorithms used by firms. It thus faces a *regulation problem*; it can restrict which attributes decision-makers can use in their predictions, but it has no impact on the design and implementation of the actual algorithms.

Social welfare is also maximised in this situation, as discrimination is zero, because the decision-makers must disclose the decision rules. However, this is an example of input regulations. As the dynamics between input data and (biased) outcomes are not clear and often counterintuitive, this is not necessarily the best approach. Cowgill & Tucker (2019) advocate instead to regulate outputs, leaving the implementation up to the decision-makers themselves. While this approach is more concerned with actually reducing bias than complying with a fair process, the implementation is conveniently left to the private sector.

In reality, however, both approaches face several constraints. First, if there is regulation in place that makes algorithmic audits mandatory, then these audits need to be performed by someone. Which party is responsible for the audits is not yet clear. Some researchers interpret the 'right to explanation' embedded within the GDPR as mandating algorithmic audits (Edwards & Veale, 2017). In Europe, this would place the responsibility of performing audits on, for instance, the national Data Protection Agencies (DPAs) (Casey et al., 2019). The GDPR also enhances the enforcement powers of the DPAs, which makes them a viable option as the responsible power.

Secondly, Rambachan et al. (2020) define an algorithmic audit as the full disclosure of the data, trained predictive algorithm and the corresponding decision rules. This would lead to an optimal social welfare level, in which bias is eradicated. However, current state-of-the-art algorithmic audit techniques consist of specialized software tools that take a predictive algorithm, training data and/or predictions made by the algorithm and test these against (a combination of various) fairness definitions. As stated in Section 2.3, certain fairness definitions are mutually incompatible. None of them are able to guarantee a complete lack of bias, as a predictive algorithm deemed fair by one definition may contain a bias according to another fairness definition.

2.4.2 Fairness accuracy trade-off and the social welfare function

Many researchers have shown there is a trade-off between achieving different fairness measures and accuracy in predictive modelling, as methods for fair machine learning place additional constraints or a penalization term on the learning algorithms. This in turn limits the information available to the algorithm. See, for instance, Liu & Vicente (2020) and (Berk et al., 2017) for empirical

evaluations of this trade-off for multiple definitions of fairness and fairness interventions. Thus, making algorithms fairer comes at the cost of predictive accuracy.

In traditional economic approaches, this sort of trade-off is encoded in a social welfare function. Social welfare functions were introduced by Bergson in 1938 in order to compare social alternatives. A social welfare function aggregates the preferences of each individual in a society, in order to help guide decisions that influence social welfare (Weymark, 2016).

Rambachan et al. (2020) apply this framework to regulating fairness in algorithms. In their work, social welfare is defined as society's preferences over the outcomes of a decision-making process. Their social welfare function tries to define a societally optimal outcome for this regulation process, by defining the function in terms of the *outcome of interest* used in creating the algorithm. The social welfare function is then defined as the weighted average outcome of interest among individuals that receive a 'positive' decision by the predictive algorithm. So, again continuing with the same examples, it is optimal, for instance, when the total predicted productivity of hired candidates is maximized; or when the total recidivism rate for released defendants is as low as possible; or when the credit scores of clients with approved loans is as high as possible. The social welfare function contains the *weighted* average of these outcomes, with the weights giving the possibility to express a preference over certain outcomes for specific groups. For instance, if society views female candidates as being historically disadvantaged in job applications, then the weight of this group can be increased. As hiring a female candidate would then increase social welfare compared to hiring a male candidate (assuming the rest of their characteristics are exactly identical), decision-makers would be expected to hire the female candidate. This mechanism allows the social welfare function to give a preference to more equitable outcomes, which are assumed to be societally desirable.

2.5 Overview

As discussed above, research in algorithmic bias spans multiple fields. Many researchers have focused on different interpretations of what constitutes a fair algorithm from a computational or mathematical perspective. They have shown that it is impossible to satisfy some of these notions of fairness simultaneously. Therefore, when utilising these group fairness notions to determine whether using an algorithm leads to disparate impact for different groups, an algorithm will never be deemed completely free of bias according to all fairness definitions. Increasingly, researchers are seeking to add an economic perspective to the discussion on algorithmic fairness (Cowgill & Tucker,

2019; Rambachan et al., 2020). They posit that the context within which an algorithm operates, and the way it is used, is as important for fairness as its technical specifications.

In the case of the COMPAS risk tool, its defenders declared that the tool was fair as the algorithm made mistakes for black and white defendants at roughly the same rate; the algorithmic was deemed fair according to the definition of (Dieterich et al., 2016). The investigation by ProPublica, however, assessed how fair the outcomes of the algorithm are using entirely different measures; they found that black defendants were more likely to be wrongfully predicted high-risk when they were actually low risk (higher False Positive Rate). White defendants, on the other hand, were more likely to be wrongfully predicted low-risk when they were actually high-risk (higher False Negative Rate). This shows that the debate around algorithmic fairness should not just be centred on the technical aspects, but that the specific use case of an algorithmic also has a role in determining what constitutes fair outcomes.

Likewise, how governments should regulate algorithms to reduce (the impact of) bias on the world around us is also hotly debated. An integral part of many solutions seems to be the algorithmic audit. Such an audit would ideally reveal the bias present in or exacerbated by an algorithm. Once this bias has been revealed, various measures can be utilised to mitigate this bias.

Rambachan et al. (2020) researched regulation regarding fairness in algorithms. They used a social welfare function to model societal preferences over the outcome of algorithmic decision-making processes in order to find optimal regulation of algorithms. Their work, however, is confined to finding bias in algorithms, and does not consider recent work on bias mitigation techniques, which aims to combat the biases found in algorithms.

Research has shown that there is a trade-off between fairness and predictive accuracy in predictive modelling, as measures that enhance fairness usually limit the information available to the model. The social welfare function used in Rambachan et al. (2020) can be conceptualized as a fairness measure, albeit one that does try to combine payoffs from diversity, equity and efficiency (Cowgill & Tucker, 2019). Whether this trade-off also exists for social welfare functions has to my knowledge not yet been investigated. This research will therefore try to fill that gap. **SQ5:** *Is there a trade-off between social welfare and prediction accuracy in the context of bias mitigation techniques?* will be answered in the following section, Section 3: Method. In order to give more structure to that section, I will test the following hypothesis:

H1: Treating an algorithm with a fairness intervention has a positive effect on social welfare

If an algorithm receives a fairness intervention, then the algorithm should become more fair, and thus, according to Rambachan et al. (2020), social welfare should increase. This makes intuitive sense, as it should be societally desirable to produce fairer algorithms and the social welfare function should reflect this societal desirability.

H2: Treating an algorithm with a fairness intervention has a negative effect on predictive accuracy

Previous studies such as Liu & Vicente (2020) and Friedler et al. (2019) have supported H2. Taken together, if my research shows support for both H1 and H2, then that will confirm the existence of a trade-off between social welfare score and predictive accuracy.

Section 3: Method

The previous chapter offered insights to the state-of-the-art research on bias in algorithmic decision-making and how to regulate this. It also described where the literature is lacking; specifically, in quantifying the social desirability of implementing bias mitigation techniques to combat bias in algorithms, and whether there is a trade-off between this social desirability and predictive accuracy. In order to analyse this trade-off, to answer **SQ4:** *Is there a trade-off between social welfare and prediction accuracy in the context of bias mitigation techniques?*, the following method is proposed.

3.1 Creating biased synthetic datasets

While it is true biased data is not needed in order to generate a biased model, in this approach, generating biased data synthetically allows for a measure of control of bias present in the models. First, an introduction on formalizing notions of fairness, in order to explain how audit tool test for fairness.

3.1.1 Formalizing fairness notions

In Section 2.2.1, the theoretical foundations of many differing fairness definitions were discussed. In this section, the definitions that will be used for this research will be elaborated on, and their implementation will be discussed. Note that the description of the various fairness definitions includes the terms *positive* and *negative* labels or predictions. These refer to the labels or predictions having the value 1 (positive) or 0 (negative). As the data used here are generated synthetically, positive and negative have no actual meaning, but can be seen as useful constructs in understanding the various fairness definitions.

As algorithmic audit tools aggregate and present results on bias on the dataset as a whole, it is natural to consider group fairness measures to be most applicable. The tools simply lack the features to test for more complicated definitions such as individual or causal fairness.

3.1.1.1 Statistical metrics used

Group fairness definitions are based on a number of metrics that are commonly used throughout machine learning. They can be summarized using a confusion matrix; a table that is used to describe the accuracy of a classification model.

The rows of the matrix refer to the predicted classes and the column to actual classes. See Table 1. In the case of this research, these classes can be positive or negative (1 or 0). The following concepts are based on this confusion matrix and can be used to determine fairness.

True positive (TP): the predicted and the actual outcome are both in the positive class.

False positive (FP): the predicted outcome is in the positive class, but the actual outcome is in the negative class.

True negative (TN): the predicted and actual outcome are both in the negative class.

False negative (FN): the predicted outcome is in the negative class, but the actual outcome is in the positive class.

True positive rate (TPR): the fraction of positive cases that is predicted to be positive out of all actual positive cases.

False positive rate (FPR): the fraction of negative cases that is predicted to be positive out of all actual negative cases.

True negative rate (TNR): the fraction of negative cases that is predicted to be positive out of all actual negative cases.

False negative rate (FNR): the fraction of positive cases that is predicted to be negative out of all actual positive cases.

Positive predictive value (PPV): the fraction of positive cases that is predicted to be positive out of all predicted positive cases.

False discovery rate (FDR): the fraction of negative cases that is predicted to be positive out of all predicted positive cases.

	Actual - Positive	Actual - Negative
Predicted - Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR $\frac{FP}{FP+TN}$
Predicted - Negative	**False Negative (FN)** FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** TNR = $\frac{TN}{TN+FP}$

Table 1: Confusion matrix. Note: Adapted from Verma & Rubin (2018)

3.1.2 Creating biased datasets

The method for creating synthetically biased datasets is adapted from the one used by Zafar et al. (2017). These dataset consist of *n* instances (number of rows). Zafar et al. (2017) set the value of *n* at 10 000. However, due to available processing power owing to Google Colab, the value of *n* has been chosen as 500 000, allowing for greater accuracy while still being computational feasible. This dataset can be conceptualized as consisting of 500 000 individuals. Each of these individuals has an *outcome of interest*, which is what the model will try to predict. This outcome of interest is called Y, and is a binary variable that is drawn from a discrete uniform distribution $Y \sim U(0,1)$. Every individual also has a *sensitive attribute*, S, which is also a binary variable drawn from a discrete uniform distribution $S \sim U(0,1)$. This sensitive attribute can represent a binary protected attribute such as gender. The individuals in the dataset are grouped using this sensitive attribute, so when this report mentions a *group*, it refers to a group of individuals sharing the same sensitive attribute value. In the context of fairness, there is usually one group who is referred to as the *privileged group*, with one or more groups then being the *unprivileged group*. This privileged group is the group that in one way or another receives better treatment, better outcomes or enjoys better fairness metrics.

Each row also has a two-dimensional user feature vector, called x. This feature vector can be thought of as the characteristics describing this individual. These features are varied in three ways to create three distinctly biased datasets.

3.1.2.1 Different False Positive Rates

In order to ensure that the two groups have different False Positive Rates, the user feature vector x is sampled from the following distributions.

$$p(x|S=0,Y=1) = N([2,2],[3,1;1,3])$$
$$p(x|S=1,Y=1) = N([2,2],[3,1;1,3])$$
$$p(x|S=0,Y=0) = N([1,1],[3,3;1,3])$$
$$p(x|S=1,Y=0) = N([-2,-2],[3,1;1,3])$$

As the sensitive attribute S and outcome of interest Y are both uniformly distributed, and the number of samples drawn is quite large, it can be expected that each distribution is represented in ¼ of the dataset. This ensures that the two groups have different distributions for the negative classes; Zafar et al. (2017) call this fairness metric *disparate mistreatment on FPR.*

3.1.2.2 Different False Negative Rates

In order to ensure that the two groups have different False Negative Rates, the user feature vector x is sampled from the following distributions.

$$p(x|S=0,Y=1) = N([1,1],[3,3;1,3])$$
$$p(x|S=1,Y=1) = N([-2,-2],[3,1;1,3])$$
$$p(x|S=0,Y=0) = N([2,2],[3,1;1,3])$$
$$p(x|S=1,Y=0) = N([2,2],[3,1;1,3])$$

As the sensitive attribute S and outcome of interest Y are both uniformly distributed, and the number of samples drawn is quite large, it can be expected that each distribution is represented in ¼ of the dataset. This ensures that the two groups have different distributions for the positive classes. Zafar et al. (2017) call this fairness metric *disparate mistreatment on FNR.*

3.1.2.2 Different False Positive Rates and different False Negative Rates

In the case where the groups have both different False Negative Rates as well as False Positive Rates, two scenarios are tested. The first scenario is where the differences in False Negative Rates and False Positive Rates between the two groups have the same sign, i.e. both False Negative Rates and False Positive Rates are higher for one group then for the other. This scenario can arise when one group is harder to classify. The user feature vector x is sampled from the following distributions to simulate this.

$$p(x|S=0,Y=1) = N([2,0],[5,1;1,5])$$
$$p(x|S=1,Y=1) = N([2,3],[5,1;1,5])$$

$$p(x|S=0,Y=0) = N([-1,-1],[5{,}1;1{,}5])$$
$$p(x|S=1,Y=0) = N([-1{,}0],[5{,}1;1{,}5])$$

In the other scenario, the differences in False Negative Rates and False Positive Rates between the two groups have the opposite sign, i.e. the False Negative Rate is lower for one group, while the False Positive Rates are higher than the other. This can be the case when the model disproportionally favours individuals from the privileged group when they are in the positive class (have Y = 1), while at the same time disproportionally disfavouring individuals from the unprivileged group when they are in the negative class (have Y = 0). The user feature vector x is sampled from the following distributions to simulate this.

$$p(x|S=0,Y=1) = N([1{,}2],[5{,}2;2{,}5])$$
$$p(x|S=1,Y=1) = N([2{,}3],[10{,}1;1{,}4])$$
$$p(x|S=0,Y=0) = N([0,-1],[7{,}1;1{,}7])$$
$$p(x|S=1,Y=0) = N([-5{,}0],[5{,}1;1{,}5])$$

Zafar et al. (2017) call this fairness metric *disparate mistreatment on both FPR and FNR*, but it is more commonly known in the literature as *equalized odds* (Gajane & Pechenizkiy, 2017; Verma & Rubin, 2018).

3.2 Training classifiers on datasets

Next, each dataset is split into a training and a test set, with a split of 80-20. For each dataset, a logistic regression model is trained on x (training) to predict Y (training). Logistic regression is a linear regression model that models the probabilities of two possible outcomes. It is also the type of model under inspection in Zafar et al. (2017). As it models a linear relationship, if the instances containing 0 (negative) and 1 (positive) outcomes are not linearly separable, the model is bound to misclassify a portion of the instances. The predicted outputs of these models, denoted here as $\hat{Y}$, will then be used in testing the audit tools.

3.3 Treating the biased classifiers/datasets with bias mitigation techniques

In order to combat the bias found in the trained classifier, a number of bias mitigation techniques are used. To streamline the process of implementing these bias mitigation techniques, the audit tool

AI Fairness 360 (AIF360) by IBM is used. AIF360 is a comprehensive algorithmic audit tools, which can detect violations of group fairness definitions in trained classifiers. It is then also able to mitigate these biases, by modifying either the training data, the algorithm or the predictions themselves. It supports eleven fairness intervention, but only a subset of these are used in this research. Most method include some parameter that can be tweaked. Unless stated otherwise, these parameters were left to their default values as much as possible in order to minimise the number dimensions tested.

3.3.1 Preprocessing methods

3.3.1.1 Disparate Impact Remover

Disparate Impact Remover was introduced by Feldman et al. (2014) and is a preprocessing technique that edits the *feature* values (in this report, the values of x). It aims to modify the marginal distributions of these features so that subsets of that attribute are equal for the different sensitive attribute groups (Friedler et al., 2019). This method uses a parameter called *repair_level* to control the trade-off between fairness and accuracy, where *repair_level* = 0 indicates that no fairness considerations and *repair_level* = 1 maximises fairness (Pessach & Shmueli, 2020). Then, similar to the method for the untreated data above, this modified dataset is split into a train and test set with a split of 80-20 and a logistic classifier is trained on the training data.

3.3.1.2 Reweighing

Reweighing is a preprocessing technique introduced in Kamiran & Calders (2012) that gives a weight to combination of Y and S in order to increase fairness. Exactly which fairness metric is considered is not made explicit. The weights are then added as an extra feature in x, after which this dataset is split into a train and test set with a split of 80-20 and a logistic classifier is trained on the training data.

3.3.1.3 Learning Fair Representations

Learning Fair Representations is a technique introduced by Zemel et al. (2013) that tries to achieve equal Positive Predictive Values and individual fairness simultaneously by learning a representation of the data that obfuscates information on the sensitive attribute. However, using this algorithm to transform the data used in this report leads to a perfect model. A perfect model predicts every data point without error. If a model does not have any errors, it also does not give different predictions for different sensitive attribute groups, so it does not make sense to compute the social welfare score, as this will always be 1. Therefore, this technique is not used in this research.

3.3.1.4 Optimized Preprocessing

Optimized Preprocessing was introduced by Calmon, Wei, Vinzamuri, Ramamurthy, & Varshney (2017) that edits both the features in x as well as the outcome of interest Y. However, this algorithm uses a distortion constraint to account for individual fairness; this research is only concerned with group fairness, as determining this distortion constraint involves setting a cost for unfair classification, which is outside of the scope of this research. Therefore, it is not used.

3.3.2 Inprocesssing methods

3.3.2.1 Prejudice Remover

Kamishima, Akaho, Asoh, & Sakuma (2012) introduced Prejudice Remover, an algorithm that adds a regularization term to the standard log-likelihood loss function of a classifier that penalizes discrimination over sensitive attribute groups. Like in *L1* and *L2* regularization, the Prejudice Remover Algorithm employs a hyperparameter that controls how severe this penalization is (D'Alessandro et al., 2017).

3.3.2.2 Meta Algorithm for Fair Classification

This technique was introduced by Celis, Huang, Keswani, & Vishnoi (2019) and is similar to Prejudice Remover in that it adds a fairness constraint to the learning function. It is able to optimize for either False Discovery Rates or Statistical Parity. For this research, only the standard parameter False Discovery Rate was used. Using this algorithm on the dataset that is biased on FNR leads to excessive use of computational power, resulting in out-of-memory issues. Therefore, this algorithm is not used for that specific dataset.

3.3.2.3 Adversarial Debiasing

Adversarial Debiasing is a sophisticated debiasing method introduced by Zhang, Lemoine, & Mitchell (2018) that leverages adversarial learning in order to achieve equality of odds. It can be extended to accommodate other fairness definitions as well as regression tasks, but sadly, in AIF360, it is implemented in Tensorflow 1. As Google Colab is used for this research, which only supports Tensorflow 2.0 and higher, that the syntax does not work anymore. As reimplementing this package in Tensorflow 2.0 is outside the scope of this project, this method is not used.

3.3.2.4 Adversarial-Robustness-Toolbox

The Adversarial-Robustness-Toolbox is a Python library for Machine Learning security, in order to defend models from a number of malicious attacks (Nicolae et al., 2018). As their goal is different from the goal of this research, it will not be implemented here.

3.3.2.5 Rich Subgroup Fairness

Rich Subgroup Fairness is an inprocessing method introduced in Kearns, Neel, Roth, & Wu (2018). This algorithm is concerned with calculating fairness metrics over subgroups; combinations of different sensitive attribute groups, such as young women or white men. As the research only contains one binary sensitive attribute, subgroup fairness is outside of the scope of this research, and this method will not be used.

3.3.3 Postprocessing methods

3.3.3.1 Calibrated Equalized Odds and Equalized Odds Postprocessing

Calibrated Equalized Odds and Equalized Odds Postprocessing are listed as two separate methods in the AIF360 documentation, but they are both based on Pleiss, Raghavan, Wu, Kleinberg, & Weinberger (2017) and they lead to the same outcomes when implemented. Therefore, they will be treated as one method in this report. This method solves a linear program in order to find probabilities that it then uses to change the model output in order to optimize equal True Positive Rates and False Positive Rates for both groups. It thus uses the output from the logistic regression algorithm described in Section 3.2 and changes the outputted predictions to achieve equalized odds.

3.3.3.2 Reject Option Classification

Reject Option Classification is a postprocessing technique that takes a confidence bound around the decision boundary created by a model and switches negative predictions to positive predictions for the unprivileged group and vice versa for the privileged group around this confidence bound (Kamiran, Karim, & Zhang, 2012). It is able to find the best confidence bound by itself, but it is a very computationally expensive method, with a large number of parameters: the smallest and highest classification thresholds used in the optimization process; the number of classification thresholds used in the optimization search; the number of margins it should use in the search; the fairness metric used for optimization (it supports Statistical Parity, Equalized Odds and Equality of Opportunity); and the upper and lower bound of the constraint on the fairness metric value.

3.4 Evaluating the classifiers/datasets

For each dataset, a number of metrics are calculated before and after treatment with fairness interventions. For every dataset, the mean and standard deviations of the outcome of interest, the sensitive attribute and x are reported, in order to see whether the data was generated successfully. These statistics are also reported for the train and test sets separately, to ensure that the split was performed randomly and the distributions over both sets remain the same. Then, after each model has been trained on the training sets, a number of performance metrics are reported. These metrics are calculated over the test sets, in order to ensure that the models are not overfitted to the training data. As the datasets generated will contain disparate impact on False Positive Rates and False Negative Rates, these metrics will be reported for both groups, both taken together and separately, as well as the difference between the groups. The accuracy (ACC) of the trained classifier will also be reported, and is calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

It will also be reported separately for each group. All these metrics will also be reported for the classifier after the fairness intervention. According to *H2: Treating an algorithm with a fairness intervention has a negative effect on predictive accuracy*, ACC is expected to be lower after fairness interventions than before.

3.5 Calculating social welfare scores

Following Rambachan et al. (2020), a social welfare function is defined in this research as the average outcome of interest over the proportion of individuals that receives a positive outcome. In this research, a social welfare score is calculated over each dataset, before and after it receives a fairness intervention. It is thus calculated as the arithmetic mean of Y of the subsample where $\hat{Y} = 1$. A perfect classifier would have a social welfare score of 1, as the predictions would perfectly align with the outcomes of interest. This metric thus penalizes errors in both positive and negative classes. According to *H1: Treating an algorithm with a fairness intervention has a positive effect on social welfare*, social welfare is thus expected to be higher after the fairness interventions than before.

3.6 Overview

The following steps will thus be followed in order to analyse whether bias mitigation techniques increase social welfare.

1. Creating biased synthetic datasets
2. Training classifiers on datasets
3. Treating the biased classifiers/datasets with bias mitigation techniques
4. Evaluating the classifiers/datasets
5. Calculating social welfare scores on the outcomes of the untreated- and treated classifiers

In this research, the algorithmic audit tool AIF360 is used to implement bias mitigation techniques. Table 2 shows which bias mitigation techniques are supported by AIF360 and whether they are used in this research.

In the case of COMPAS, the algorithm that predicted risk scores for defendants was subjected to an audit by ProPublica, in which a range of different fairness metrics were calculated to see whether the algorithm was deemed fair. However, after the report by ProPublica, Northpointe, the company behind the creation of the COMPAS tool, released its own report, criticising the approach taken by ProPublica (Dieterich et al., 2016). Both parties are able to show that the algorithm conforms to some fairness approaches, but not to others. However, missing from this discussion is a perspective on the impact the use of this algorithm has on a societal level. Different fairness measures are useful in different contexts (Green & Hu, 2018). It is not clear which metrics are most useful in the case of COMPAS. Therefore, the methodology used in this report tries to move the discussion away from the use of these fairness metrics, by directly measuring the social impact through the use of a social welfare function.

Type of intervention	Bias mitigation technique	Used in this research
Preprocessing	**Disparate Impact Remover** (Feldman et al., 2014)	**Yes**
	Reweighing (Kamiran & Calders, 2012)	**Yes**
	Learning Fair Representations (Zemel et al., 2013)	**No**
	Optimized Preprocessing (Calmon et al., 2017)	**No**
Inprocessing	**Prejudice Remover** (Kamishima et al., 2012)	**Yes**
	Meta Algorithm for Fair Classification (Celis, Huang, Keswani, & Vishnoi, 2019)	**Yes**
	Adversarial Debiasing (Zhang et al., 2018)	**No**
	Adversarial-Robustness-Toolbox (Nicolae et al., 2018)	**No**
	Rich Subgroup Fairness (Kearns et al., 2018)	**No**
Postprocessing	**Calibrated Equalized Odds** (Hardt et al., 2016; Pleiss et al., 2017)	**Yes**
	Reject Option Classification (Kamiran et al., 2012)	**Yes**

Table 2: Fairness interventions supported by AIF360

Section 4: Results

In this section, the main results of the empirical analysis are presented. In order to see whether bias mitigation techniques lead to more socially optimal outcomes, social welfare scores were calculated for a number of biased datasets before and after receiving bias mitigations.

Each dataset can be conceptualized as a pool of candidates for a screening decision, with every dataset being biased in a different way. Every row in the dataset is a single candidate with attributes *outcome_of_interest*, *sensitive_att*, x_1 and x_1. There are 500 000 candidates in the dataset, half of which belong to sensitive attribute group 0, and the other half to sensitive attribute group 1. The datasets are split into a training set and a test set with a ratio of 80/20. Please note that all metrics reported for the mitigation techniques are metrics calculated on the test set only.

4.1 Disparate mistreatment on FPR

An unconstrained logistic regression classifier was trained on this training data, using only x_1 and x_2 to predict the *outcome_of_interest*. The output of this model in term of probabilities is called *proba* and the resulting binary prediction is called decision. This classifier attains a test set accuracy of 0.77, but leads to a difference in False Positive Rate D_{FPR} = 0.53 - 0.00 = 0.53, which constitutes a clear case of disparate mistreatment in terms of false positive rates. D_{FNR} = 0.19 - 0.19 = 0. The social welfare score for this untreated dataset is 0.75.

Variable	Mean	Standard deviation
outcome_of_interest	0.50	0.50
sensitive_att	0.50	0.50
x_1	0.75	2.36
x_2	0.75	2.42
decision	0.54	0.50
proba	0.50	0.31

Table 3: Descriptive statistics for biased dataset 1 (training data).

Variable	Mean	Standard deviation
outcome_of_interest	0.50	0.50
sensitive_att	0.50	0.50
x_1	0.76	2.36
x_2	0.76	2.42
decision	0.54	0.50
proba	0.50	0.31

Table 4: Descriptive statistics for biased dataset 1 (test data).

As can be seen in Table 3 and Table 4, the characteristics for the train and test set are practically equal, as can be expected of randomly sampled data. Then, a number of fairness interventions were used in order to combat the bias present in this dataset. Where possible, the intervention was calibrated to combat the specific bias present in this dataset (disparate false positive rates). All other parameters were left to their original states, unless stated otherwise.

As seen before, the original classifier (before intervention) achieves D_{FPR} = 0.52. In Table 5 we can see that not every fairness intervention is equally adept at removing this bias.

Intervention		Accuracy	ACC group 0	ACC group 1	D_{ACC}	Change in D_{ACC}
Before intervention		**0.77**	**0.64**	**0.90**	**0.26**	
Preprocessing	Disparate Impact Remover	0.80	0.66	0.95	0.29	0.03
	Reweighing	0.77	0.64	0.90	0.26	0.00
Inprocessing	Meta Algorithm for Fair Classification	0.63	0.50	0.77	0.27	0.01
	Prejudice Remover	0.80	0.66	0.95	0.29	0.03
Postprocessing	Calibrated Equalized Odds	0.64	0.64	0.64	0.00	-0.26
	Reject Option Classification	0.77	0.62	0.93	0.30	0.04

Table 5: Accuracy before and after fairness interventions on dataset that exhibits disparate impact on FPR (test data)

Intervention		FPR	FPR group 0	FPR group 1	D_{FPR}	Change in D_{FPR}
Before intervention		**0.27**	**0.53**	**0.01**	**0.52**	
Preprocessing	Disparate Impact Remover	0.20	0.35	0.05	0.29	-0.23
	Reweighing	0.27	0.53	0.01	0.52	0.00
Inprocessing	Meta Algorithm for Fair Classification	0.73	1.00	0.46	0.54	0.02
	Prejudice Remover	0.20	0.35	0.05	0.29	-0.23
Postprocessing	Calibrated Equalized Odds	0.53	0.53	0.53	0.00	-0.52
	Reject Option Classification	0.32	0.63	0.02	0.61	0.08

Table 6: False Positive Rates before and after fairness interventions on dataset that exhibits disparate impact on FPR (test data).

4.1.1 Preprocessing methods

Disparate Impact Remover is able to remove almost half the difference in FPR by lowering the FPR of sensitive attribute group to 0.35. As this technique was created to remove any correlation between the sensitive attribute and the features (x_1 and x_1) and there is randomness involved in creating

these features, it is not surprising that the difference in FPR is not entirely erased. It improves predictive accuracy for both groups, but more so for the privileged group, resulting in a larger D_{ACC}.

The Reweighing method gives different weights to different individuals in the dataset in order to combat demographic parity, i.e. to ensure that the proportion of positive labels is equal among groups. However, since this proportion is already equal in our training data, the Reweighing method does not perform major transformations on the data; the observed differences in metrics are very small.

4.1.2 Inprocessing methods

Using Meta Algorithm for Fair Classification to classify the data leads to a classifier that performs worse in terms of accuracy as well as in terms of False Positive Rates. However, the implementation of this technique only supports training a classifier that optimizes either False Discovery Rates or statistical parity, and as such was not expected to perform well on a dataset biased on false positive rates. The method achieves an FPR of 1 for group 0, which means that all of the negative classes are predicted wrongly. As accuracy for this group is 0.50, it seems that this method outputs a positive prediction for every individual in this group.

Similar to Disparate Impact Remover, the Prejudice Remover forces a classifier to be independent from the sensitive attribute used, causing the two methods to yield almost equal results.

4.1.3 Postprocessing methods

Calibrated Equalized Odds seeks to optimize both TPRs and FPRs by changing the thresholds at which groups are classified as either positive or negative (default threshold = 0.50). Since this considers both true and false positive rates, the drop in predictive accuracy is surprising, given that the false positive rate stays at the same level. It is, however, the only method that achieves a difference in accuracy between the groups of 0, meaning that the model is equally accurate for both groups.

Reject Option Classification tries to improve fairness by favourable labels to unprivileged groups and unfavourable labels to privileged groups around a confidence band on the decision boundary, i.e. around the edge, the unprivileged group is given preference, in order to optimize for equality of opportunity (equal true positive rates). As this technique also optimizes a metric different than FPR, it is not surprising that the false positive rates barely change. In terms of accuracy, the algorithm

slightly lowers the accuracy of the unprivileged group, and increases the accuracy for the privileged one, resulting in a larger difference between the two.

4.1.4 Social welfare

Intervention		Social welfare score	Change in social welfare score
Before intervention		**0.75**	
Preprocessing	Disparate Impact Remover	0.80	6.8%
	Reweighing	0.75	0.2%
Inprocessing	Meta Algorithm for Fair Classification	0.58	-23.0%
	Prejudice Remover	0.80	6.8%
Postprocessing	Calibrated Equalized Odds	0.60	-19.6%
	Reject Option Classification	0.73	-2.6%

Table 7: Social welfare scores before and after fairness interventions on dataset that includes disparate impact on FPR (test data).

The social welfare scores calculated over the outputs of the fairness interventions follow the same pattern as the change in D_{FPR}. Those techniques that barely or not at all make changes to either the inputs, the output or the model itself naturally also do not show any big changes in social welfare score. The interventions that did significantly change the difference in FPR (Disparate Impact Remover, Prejudice Remover) also show improvements in social welfare score.

4.2 Disparate mistreatment on FNR

For descriptive statistics on the generated dataset that exhibits a bias with respect to False Negative Rates, please see Appendix A.1.

As seen before, the original classifier (before intervention) achieves $D_{FPR} = 0.52$. In Table 6 we can see that not every fairness intervention is equally adept at removing this bias.

Intervention		Accuracy	ACC group 0	ACC group 1	D_{ACC}	Change in D_{ACC}
Before intervention		**0.81**	**0.64**	**0.90**	**0.26**	
Preprocessing	Disparate Impact Remover	0.83	0.83	0.83	0.00	-0.26
	Reweighing	0.77	0.65	0.90	0.25	0.00
Inprocessing	Meta Algorithm for Fair Classification				0.00	-0.26
	Prejudice Remover	0.81	0.66	0.95	0.29	0.03
Postprocessing	Calibrated Equalized Odds	0.64	0.64	0.64	0.00	-0.26
	Reject Option Classification	0.78	0.64	0.92	0.28	0.02

Table 8: Accuracy before and after fairness interventions on dataset that exhibits disparate impact on FNR (test data)

Intervention		FNR	FNR group 0	FNR group 1	D_{FNR}	Change in D_{FNR}
Before intervention		**0.27**	**0.53**	**0.01**	**-0.52**	
Preprocessing	Disparate Impact Remover	0.17	0.17	0.17	0.00	0.52
	Reweighing	0.27	0.52	0.01	-0.51	0.00
Inprocessing	Meta Algorithm for Fair Classification	#DIV/0!			0.00	0.52
	Prejudice Remover	0.20	0.34	0.05	-0.29	0.23
Postprocessing	Calibrated Equalized Odds	0.53	0.53	0.53	0.00	0.51
	Reject Option Classification	0.16	0.18	0.15	-0.03	0.48

Table 9: False Negative Rates before and after fairness interventions on dataset that exhibits disparate impact on FNR (test data).

4.2.1 Preprocessing methods

Disparate Impact Remover is able to achieve equal FNRs while simultaneously achieving equal Accuracy for both groups. This does come at the slight cost of lower accuracy for group 1.

The Reweighing techniques manages to decrease predictive accuracy while keeping the FNRs for both groups equal.

4.2.2 Inprocessing methods

As mentioned in the section 3.3.2, due to excessive computational demands, it was not possible to use Using Meta Algorithm for Fair Classification on this dataset.

The Prejudice Remover method is able to achieve a higher accuracy for group 1 while keeping that of group 0 equal, in order to reduce the difference in False Negative Rate.

4.2.3 Postprocessing methods

Calibrated Equalized Odds lowers the accuracy of group 1 to the level of group 0, while doing the same for False Negative Rates, to achieve parity in both those metrics.

Reject Option Classification is able to keep the accuracy for both groups roughly equal, but simultaneously decreasing the difference in FNR to almost zero, while also lowering the overall False Negative Rate.

4.2.4 Social welfare

Intervention		Social welfare score	% change
Before intervention		**0.80**	
Preprocessing	Disparate Impact Remover	0.81	1.2%
	Reweighing	0.80	0.0%
Inprocessing	Meta Algorithm for Fair Classification		-100.0%
	Prejudice Remover	0.81	1.2%
Postprocessing	Calibrated Equalized Odds	0.72	-9.8%
	Reject Option Classification	0.75	-5.6%

Table 10: Social welfare scores before and after fairness interventions on dataset that includes disparate impact on FNR (test data).

The social welfare scores for this dataset seem to follow the pattern set by the Accuracy metrics; some methods are able to keep social welfare level while lowering the difference in False Negative Rates, while others sacrifice social welfare score in order to lower D_{FNR}. Not a single method was able to make significant increases in social welfare score.

4.3 Disparate mistreatment on both FPR and FNR (different sign)

For descriptive statistics on the generated dataset that exhibits a bias with respect to both False Negative Rates and False Positive rates (different sign), please see Appendix A.2.

A number of fairness interventions were used in order to combat the bias present in this dataset. Where possible, the intervention was calibrated to combat the specific bias present in this dataset (disparate False Positive and False Negative Rates). All other parameters were left to their original states, unless stated otherwise.

As seen before, the original classifier (before intervention) achieves D_{FPR} = 0.20 and D_{FNR} = -0.19. Notice that the two have different signs.

Intervention		Accuracy	ACC group 0	ACC group 1	D_{ACC}	Change in D_{ACC}
Before intervention		**0.81**	**0.81**	**0.81**	**0.00**	
Preprocessing	Disparate Impact Remover	0.83	0.83	0.83	0.00	0.00
	Reweighing	0.81	0.81	0.81	0.00	0.00
Inprocessing	Meta Algorithm for Fair Classification	0.72	0.76	0.69	-0.06	-0.06
	Prejudice Remover	0.83	0.83	0.83	0.00	0.00
Postprocessing	Calibrated Equalized Odds	0.76	0.76	0.76	0.00	0.00
	Reject Option Classification	0.82	0.82	0.82	0.00	0.00

Table 11: Accuracy before and after fairness interventions on dataset that exhibits disparate impact on FPR and FNR (different sign) (test data)

Intervention		FPR	FPR group 0	FPR group 1	D_{FPR}	Change in D_{FPR}
Before intervention		**0.19**	**0.10**	**0.29**	**0.20**	
Preprocessing	Disparate Impact Remover	0.18	0.17	0.18	0.00	-0.20
	Reweighing	0.19	0.10	0.29	0.20	0.00
Inprocessing	Meta Algorithm for Fair Classification	0.52	0.45	0.60	0.15	-0.05
	Prejudice Remover	0.17	0.18	0.17	0.00	-0.20
Postprocessing	Calibrated Equalized Odds	0.25	0.24	0.25	0.00	-0.20
	Reject Option Classification	0.18	0.18	0.19	0.01	-0.19

Table 12: False Positive Rates before and after fairness interventions on dataset that exhibits disparate impact on FPR and FNR (different sign) (test data).

Intervention		FNR	FNR group 0	FNR group 1	D_{FNR}	Change in D_{FNR}
Before intervention		**0.19**	**0.29**	**0.09**	**-0.19**	
Preprocessing	Disparate Impact Remover	0.17	0.17	0.17	0.00	0.20
	Reweighing	0.19	0.29	0.09	-0.19	0.00
Inprocessing	Meta Algorithm for Fair Classification	0.03	0.04	0.02	-0.02	0.17
	Prejudice Remover	0.17	0.17	0.17	0.00	0.20
Postprocessing	Calibrated Equalized Odds	0.24	0.24	0.24	0.00	0.19
	Reject Option Classification	0.17	0.17	0.16	-0.01	0.19

Table 13: False Negative Rates before and after fairness interventions on dataset that exhibits disparate impact on FPR and FNR (different sign) (test data).

4.3.1 Preprocessing methods

Disparate Impact Remover is able to equalize both the FPRs and the FNRs for both groups, while at the same time increasing accuracy. It does, however, increase the FNR for group 1, and the FPR for group 0 to achieve this parity.

Reweighing does not yield any improvements for this dataset.

4.3.2 Inprocessing methods

Meta Algorithm for Fair Classification decreases accuracy for both groups but even more so for group 1. In doing so, it does achieve False Negative Rates close to 0 for both groups, but at the cost of greatly increasing False Positive Rates for both. This can be explained as the algorithm is designed to optimize False Discovery Rates, which related to the False Negative Rate. It does reduce the difference in D_{FPR}.

Again, Prejudice Remover has almost equal results as Disparate Impact Remover.

4.3.3 Postprocessing methods

Both Calibrated Equalized Odds and Reject Option Classification are able achieve equal fairness metrics in both groups. Reject Option Classification has superior results in this case, achieving higher accuracy while at the same time lowering the overall FNR and FPR.

4.3.4 Social welfare

Intervention		Social welfare score	% change
Before intervention		**0.81**	
Preprocessing	Disparate Impact Remover	0.83	2.6%
	Reweighing	0.81	0.0%
Inprocessing	Meta Algorithm for Fair Classification	0.65	-19.3%
	Prejudice Remover	0.83	2.4%
Postprocessing	Calibrated Equalized Odds	0.76	-6.3%
	Reject Option Classification	0.82	1.5%

Table 14: Social welfare scores before and after fairness interventions on dataset that exhibits disparate impact on FPR and FNR (different sign) (test data)

It is worth noting that two methods, Disparate Impact Remover and Prejudice Remover, were able to slightly improve the social welfare score, which aligns closely with the increase in accuracy that they were achieve. The other methods kept social welfare score at the same level or decreased a bit.

4.4 Disparate mistreatment on both FPR and FNR (same sign)

For descriptive statistics on the generated dataset that exhibits a bias with respect to both False Negative Rates and False Positive rates (different sign), please see Appendix A.3.

A number of fairness interventions were used in order to combat the bias present in this dataset. Where possible, the intervention was calibrated to combat the specific bias present in this dataset (disparate False Positive and False Negative Rates). All other parameters were left to their original states, unless stated otherwise.

As seen before, the original classifier (before intervention) achieves D_{FPR} = -0.25 and D_{FNR} = -0.12. Notice that the two have different signs.

Intervention		Accuracy	ACC group 0	ACC group 1	D_{ACC}	Change in D_{ACC}
Before intervention		**0.82**	**0.73**	**0.92**	**0.18**	
Preprocessing	Disparate Impact Remover	0.81	0.72	0.90	0.17	-0.01
	Reweighing	0.83	0.74	0.91	0.18	-0.01
Inprocessing	Meta Algorithm for Fair Classification	0.68	0.57	0.78	0.21	0.03
	Prejudice Remover	0.84	0.74	0.93	0.19	0.00
Postprocessing	Calibrated Equalized Odds	0.73	0.73	0.73	0.00	-0.19
	Reject Option Classification	0.82	0.73	0.91	0.19	0.00

Table 15: Accuracy before and after fairness interventions on dataset that exhibits disparate impact on FPR and FNR (same sign) (test data)

Intervention		FPR	FPR group 0	FPR group 1	D_{FPR}	Change in D_{FPR}
Before intervention		**0.18**	**0.31**	**0.06**	**-0.25**	
Preprocessing	Disparate Impact Remover	0.20	0.29	0.11	-0.18	0.07
	Reweighing	0.18	0.30	0.06	-0.24	0.00
Inprocessing	Meta Algorithm for Fair Classification	0.64	0.85	0.43	-0.42	-0.18
	Prejudice Remover	0.17	0.27	0.07	-0.20	0.04
Postprocessing	Calibrated Equalized Odds	0.31	0.31	0.31	0.00	0.25
	Reject Option Classification	0.23	0.39	0.06	-0.33	-0.09

Table 16: False Positive Rates before and after fairness interventions on dataset that exhibits disparate impact on FPR and FNR (same sign) (test data)

Intervention		FNR	FNR group 0	FNR group 1	D_{FNR}	Change in D_{FNR}
Before intervention		**0.17**	**0.23**	**0.11**	**-0.12**	
Preprocessing	Disparate Impact Remover	0.19	0.27	0.10	-0.17	-0.06
	Reweighing	0.17	0.22	0.11	-0.12	0.00
Inprocessing	Meta Algorithm for Fair Classification	0.01	0.01	0.01	0.00	0.12
	Prejudice Remover	0.16	0.24	0.07	-0.17	-0.05
Postprocessing	Calibrated Equalized Odds	0.23	0.23	0.23	0.01	0.12
	Reject Option Classification	0.13	0.15	0.11	-0.04	0.08

Table 17: False Negative Rates before and after fairness interventions on dataset that exhibits disparate impact on FPR and FNR (same sign) (test data)

4.4.1 Preprocessing methods

In contrast to the previous dataset, Disparate Impact Remover is not able to equalize the FPRs and the FNRs for both groups. It does sacrifice a little accuracy in order to reduce the difference in False Positive Rates, but it simultaneously increases the difference in False Negative Rates.

Again, Reweighing does not yield any improvements for this dataset.

4.4.2 Inprocessing methods

Again, Meta Algorithm for Fair Classification decreases accuracy for both groups but even more so for group 1. In doing so, it does achieve False Negative Rates close to 0 for both groups, but at the cost of greatly increasing False Positive Rates for both. This can be explained as the algorithm is designed to optimize False Discovery Rates, which related to the False Negative Rate. This time it also increases the difference in D_{FPR}.

For this dataset, Prejudice Remover is able to increase accuracy slightly, while slightly lowering D_{FPR} and slightly increasing D_{FNR} and also slightly lowering the overall levels of FPR and FNR.

4.4.3 Postprocessing methods

For this dataset, Calibrated Equalized Odds is again able achieve equal fairness metrics in both groups, by decreasing accuracy, FPR and FNR of the privileged group to the level of the lowest group.

Reject Option Classification is able to decrease D_{FNR} while keeping Accuracy steady, albeit at the cost of an increase in D_{FPR}.

4.4.4 Social welfare

Intervention		Social welfare score	% change
Before intervention		**0.82**	
Preprocessing	Disparate Impact Remover	0.80	-1.9%
	Reweighing	0.82	0.0%
Inprocessing	Meta Algorithm for Fair Classification	0.61	-25.8%
	Prejudice Remover	0.83	1.6%
Postprocessing	Calibrated Equalized Odds	0.72	-12.8%
	Reject Option Classification	0.79	-3.2%

Table 18: Social welfare scores before and after fairness interventions on dataset that exhibits disparate impact on FPR and FNR (same sign) (test data)

For this dataset, again, social welfare scores followed the same pattern as accuracy.

4.5 Trade-offs

The goal of this Section was to find an answer to **SQ4:** *Is there a trade-off between social welfare and prediction accuracy in the context of bias mitigation techniques?* Table 19 shows for what portion of the total number of bias mitigation techniques tested, accuracy and social welfare increased, decreased, or remained neutral.

		Accuracy		
		Increases	**Neutral**	**Decreases**
Social Welfare Score	**Increases**	30%	9%	0%
	Neutral	4%	4%	0%
	Decreases	0%	9%	43%

Table 19: Proportions of bias mitigation techniques tested where Accuracy and Social Welfare Scores increase, remain neutral, or decrease.

Increases are defined in this case as any positive change larger than 0.01, and similarly for decreases. If the absolute change in metric is smaller than 0.01, than it is counted as a neutral. See Appendix B.1 for an overview of how each technique was counted. In order to show there is a trade-off between social welfare score and Accuracy, a large portion of cases should show opposing signs. However, as can be seen in Table 19, 30% of all interventions show increases in both metrics, 4% show no changes, and 43% show decreases in both metrics. This can be interpreted as a sign that

there is no trade-off between social welfare score and predictive accuracy, answering **SQ4:** *Is there a trade-off between social welfare and prediction accuracy in the context of bias mitigation techniques?*

Section 5: Conclusion, discussion and policy implications

The aim of this research was to analyse whether using bias mitigation techniques on biased datasets led to more societally desirable outcomes. Four sub-questions were asked to accomplish this. The literature review in Section 2 answered sub-questions 1 through 3, by creating insights into what constitutes algorithmic bias, how this bias can be detected and corrected for in algorithms and how a social welfare function can be used to quantify society's preferences over the outcomes of an algorithmic decision-making process. Section 3 introduced a method for analysing bias mitigation techniques through the lens of social welfare functions by generating biased synthetic datasets. The results of this analysis were presented in Section 4. This section summarizes the outcomes found in the previous sections and strives to answer the main research question, **To what extent do fairness interventions influence the trade-off between social welfare and predictive accuracy in the context of algorithmic decision-making?**

5.1 Conclusion

Rambachan et al. (2020) showed that a welfare-economics approach to algorithmic bias can produce policy implications that are distinct from the computer science approach that is usually taken by practitioners. The analysis in this report tries to further this argument by analysing bias mitigations through the lens of welfare economics.

Using predictive algorithms to aid in screening decision-making processes opens these processes up to algorithmic bias, which leads to a sub-optimal allocation of potential candidates. This is both socially and economically undesirable; both candidates and decision-makers receive benefits from optimal screening decisions. This bias can arise in multiple stages in the modelling pipeline; in the data, in the predictive model itself and in the way the model is used. Modern algorithmic audit tools are capable of exposing these biases by testing the model outputs against different fairness definitions. Moreover, some tools are capable of reducing or removing bias by means of bias mitigation techniques.

This project researched this dynamic by creating biased datasets, subjecting them to bias mitigation techniques and analysing them using a social welfare function. By examining the social welfare scores before and after treating each biased model/dataset with the chosen fairness intervention, it is possible to see whether this treatment results in more desirable outcomes, both socially and economically. This, in turn, creates insights for the regulation of algorithms.

The results from the analysis in Section 4 found there was not enough evidence to support the theory that there is indeed a trade-off between social welfare score, which measure social desirability, and predictive accuracy. Thus, the answer to the main research question, **To what extent do fairness interventions influence the trade-off between social welfare and predictive accuracy in the context of algorithmic decision-making?** is that no empirical evidence to suggest there is a trade-off present between social welfare score and predictive accuracy. Taken at face value, this result would suggest that continuing the push for ever greater predictive accuracy would in the long-term result in the most societally desirable outcome. However, this is not in line with recent research that suggests there is a trade-off between predictive accuracy and fairness; some predictive accuracy has to be sacrificed in order to enable algorithms to lead to more equitable outcomes for all groups affected by the algorithm (Berk et al., 2017; Liu & Vicente, 2020).

5.2 Theoretical contributions

This research contributed to existing research in several ways. First, an overview of the existing literature on algorithmic bias in the context of screening decisions, and how to combat this bias, was created. Existing research focused mainly on mathematical notions of fairness, but this research added a welfare economics perspective to the discussion on combating bias in algorithms. Secondly, different fairness interventions are theoretically embedded in the context of social welfare. Thirdly, the method for creating biased synthetic biased datasets from Zafar et al. (2017) was used to empirically test a number of different bias mitigation techniques. Fourthly, the social welfare function as used in Rambachan et al. (2020) was implemented empirically and used to determine social desirability of a number of different bias mitigation techniques.

5.3 Policy implications

The results showed that almost half of the bias mitigations tested led to both decreased social welfare scores, as well as decreased accuracy scores. While this would imply that bias mitigation techniques do more harm than good, these results are not in line with current research. While the social welfare score as used in this research might not be the most accurate approximation of what is actually societally desirable, the results did show that many fairness interventions were able to lower the difference in different metrics for both groups, often at a small cost of predictive accuracy. This means that fairness interventions are effective at removing bias of different sorts. While they might not completely make algorithms free of bias, they are often able to at least make sure some metrics are equal for all groups, so that the algorithms make the same mistakes for all groups. Section 2.5 mentioned that the issue that Larson et al. (2016) had with the COMPAS case, was that black defendants had a higher False Positive Rate and a lower False Negative Rate than white

defendants. Continuing this case, the above result would imply that, had the COMPAS tool been subjected to a bias mitigation algorithms, it would likely have resulted in increasing the False Positive Rate for white defendants and increasing False Negative Rate for black defendants, therefore increasing the error rates, and thus decreasing predictive accuracy, for both groups. While this would have led to more equitable outcomes, as defendants from all races would face the risk of being incarcerated while being innocent or vice versa, it would also lead to more overall errors, as predictive accuracy is lowered. This would lead to more innocent people in prison and more dangerous people on the street, hence the decrease in social welfare score. However, these errors would be distributed more equitably. With the current algorithm, the innocent people are prison are disproportionately more often black and the dangerous people not incarcerated are disproportionately more often white. Which situation is preferable, is a matter of public debate.

Also, the AIF360 tool is very accessible for practitioners, as well as free to use. These factors make this tool a viable option for decision-makers using algorithms to augment their decision-making process for auditing their algorithms and using the various bias mitigation techniques to limit the disparate impact algorithms can have. This research has shown that, while not functioning perfectly, state-of-the-art algorithmic audit tools are already able to remove some bias. While this often comes at a small cost of accuracy, some audit tools are able to even augment predictive accuracy, while at the same time making the algorithm fairer. Thus, I can conclude that policy makers should regard algorithmic audit tools and bias mitigation techniques as viable tools in regulating algorithmic decision-making. It should be noted that the field of fairness contains many different metrics on which to test outcomes, and while AIF360 tries to implement as many of them as possible, it does not give guidance on what metric or algorithm is useful in which context.

As the tools do need access to the ground truth labels and the predictions outputted by the predictive algorithms, these audit tools could best be used by the parties developing the predictive algorithms themselves. However, in order to provide independent third-party auditing, a number of different options arise. Some researchers interpret the ‘right to explanation’ embedded within the GDPR as mandating algorithmic audits (Edwards & Veale, 2017). This would place the responsibility of performing audits on the national Data Protection Agencies (in Europe) (Casey et al., 2019). The GDPR also enhances the enforcement powers of the DPAs, which makes them a viable option as the responsible power. This also fits within the vision of government jobs of the future by Deloitte (Egger, Datar, & Coltin, 2019). An alternative could be found in external commercial parties, as is it performed now by a select few companies such as O'Neil Risk Consulting & Algorithmic Auditing and Eticas (Winick, 2018).

Also, GDPR places constraints on the personal attributes that can be used by algorithms. However, to be able to audit algorithms for possible biases with tools such as AIF360, these personal attributes need to be included in datasets. For instance, it is impossible to calculate accuracy rates for males and females if a dataset does not contain a variable sex. This issue will become more problematic as algorithmic audits become more widespread. Furthermore, the additional accountability of algorithmic decision-making systems required under the GDPR implies that algorithms will someday be required to be completely free of bias. The results in this research suggest that at least in some cases, this will result in worse predictive accuracy, and thus to algorithms becoming more error prone. Unless researchers find techniques that are able to debias algorithms with lower penalties predictive accuracy, this will lead to decision-makers having to either rely less on predictive algorithms or seek other ways to make accurate predictions. This can for instance take the form of gathering more data (on all population subgroups) on which to train algorithms.

5.4 Limitations and future research

Even though theoretical and policy implications can be drawn from the results of the analysis performed, this research has made some assumptions that can be used as avenues for further research. These will be discussed in this section, in no particular order.

- *Data used*: Although this research has shown that synthetically generated data can be a useful tool to test the capabilities of bias mitigation techniques, it is not able to capture the complexities that real data can offer. There are many different datasets that are specifically tailored for algorithmic fairness tasks that are freely available. Further research can test whether using the same methodology on real-life data yields different results.
- *Bias mitigation techniques*: This research only utilised the algorithms available in the AI Fairness 360 toolbox, as they were easy to deploy and using the AIF360 implementation meant that they all shared the same syntax. This research did not even use all of the algorithms available in this toolbox. See Section 3.3 for an overview of why some were not used. Many more bias mitigation techniques have been developed for group fairness in supervised learning that are not included in AIF360, such as the Reductions Approach, Discrimination Aware Tree Construction, Two Naïve Bayes and Variational Fair Autoencoder (Pessach & Shmueli, 2020). Furthermore, individual fairness is alternative approach that assumes that instead of achieving parity over fairness metrics over groups, similar individuals should be treated similarly. Future research could start by constructing a social welfare score for individuals.

- *Social welfare score*: In Rambachan et al. (2020), the social welfare function is constructed as the weighted average of the outcome of interest for individuals accepted in the screening decision. In this research, this weight has been kept constant and equal for both groups. Research suggests some bias comes from groups of people being underrepresented in the data (Suresh & Guttag, 2019). The social welfare function allows for positive discrimination, i.e. expressing a preference for underprivileged groups, by giving more weight to individuals in these groups. This was not implemented in this research but finding the right value for this weight could be a fruitful avenue for further research.
- *Fairness metrics tested*: The metrics tested in this research, disparate impact on False Positive Rates, False Negative Rates, or both, were used as datasets containing disparate impact on these metrics were developed in Zafar et al. (2017). Thus, this research only focused on fairness as defined by Equality of Opportunity. Further research could explore whether the trade-off between social welfare score and accuracy does hold for other fairness definitions, such as those mentioned in Section 2.2.1
- *Predictive models tested*: For this research, only logistic regression classifiers were trained as predictive models. Friedler et al. (2019) stress that preprocessing steps have a large influence on how well fairness metrics work on a certain algorithm or dataset. Therefore, it can be reasonably assumed that substituting the predictive algorithm for another type of algorithm can also have an influence.
- *Parameters of fairness interventions*: All of the parameters in the various fairness interventions were left to their default values, or else the computational demands for this research would become too great. In order for the bias mitigations to realize their true potential, these parameters should be tweaked.

Despite these limitations, this findings in this research can hopefully still be used as a starting point for further research.

References

Ahn, Y., & Lin, Y.-R. (2019). FairSight: Visual Analytics for Fairness in Decision Making. *IEEE Transactions on Visualization and Computer Graphics*, 1–1. https://doi.org/10.1109/tvcg.2019.2934262

Alexander, L. (1992). What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes, and Proxies. *University of Pennsylvania Law Review*, *141*(1), 149. https://doi.org/10.2307/3312397

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2477899

Belinkov, Y., & Glass, J. (2018). *Analysis Methods in Neural Language Processing: A Survey*. Retrieved from http://arxiv.org/abs/1812.08951

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., … Zhang, Y. (2019). AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. *IBM Journal of Research and Development*, 1–1. https://doi.org/10.1147/jrd.2019.2942287

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., … Roth, A. (2017). A Convex Framework for Fair Regression. *ArXiv*. Retrieved from http://arxiv.org/abs/1706.02409

Cabrera, Á. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D. H. (2019). *FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning*. Retrieved from http://arxiv.org/abs/1904.05419

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems* (Vol. 30).

Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise. *Berkeley Technology Law Journal*, *34*(1), 143. https://doi.org/10.15779/Z38M32N986

Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 319–328.

https://doi.org/10.1145/3287560.3287586

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*. https://doi.org/10.1145/3173574.3174225

Chouldechova, A. (2017). *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*. Retrieved from http://arxiv.org/abs/1703.00056

Chouldechova, A., Putnam-Hornstein, E., Dworak-Peck, S., Benavides-Prado, D., Fialko, O., Vaithianathan, R., … Wilson, C. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research*, *81*, 1–15. Retrieved from http://proceedings.mlr.press/v81/chouldechova18a/chouldechova18a.pdf

Cowgill, B., & Tucker, C. E. (2019). Economics, Fairness and Algorithmic Bias. *SSRN Electronic Journal*, 1–65. https://doi.org/10.2139/ssrn.3361280

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. https://doi.org/10.1089/big.2016.0048

Datenethikkommission. (2019). *Opinion of the Data Ethics Commission - Executive Summary*.

Dieterich, W., Mendoza, C., & Tim Brennan, M. (2016). *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S. J., O'Brien, D., … Wood, A. (2017). Accountability of AI Under the Law: The Role of Explanation. *Ssrn*, 1–15. https://doi.org/10.2139/ssrn.3064761

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1). https://doi.org/10.1126/sciadv.aao5580

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Edelman, B. (2011). Bias in Search Results: Diagnosis and Response. *Indian Journal of Law and Technology*, *7*, 16–32. https://doi.org/10.1525/sp.2007.54.1.23.

Egger, W. D., Datar, A., & Coltin, K. (2019). Government Jobs of the Future. Retrieved from Deloitte Insights website: https://www2.deloitte.com/content/dam/insights/us/articles/4767_FoW-in-govt/DI_Algorithm-auditor.pdf

Elisa Celis, L., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 319–328. https://doi.org/10.1145/3287560.3287586

Eslami, M., Aleyasen, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). FeedVis: A path for exploring news feed curation algorithms. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, *2015-Janua*, 65–68. https://doi.org/10.1145/2685553.2702690

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., … Sandvig, C. (2015). I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. *Conference on Human Factors in Computing Systems - Proceedings*, *2015-April*, 153–162. https://doi.org/10.1145/2702123.2702556

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

European Parliament, & Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da*. , (2016).

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2014). *Certifying and removing disparate impact*. Retrieved from http://arxiv.org/abs/1412.3756

Friedler, S. A., Choudhary, S., Scheidegger, C., Hamilton, E. P., Venkatasubramanian, S., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 329–338. https://doi.org/10.1145/3287560.3287589

Gajane, P., & Pechenizkiy, M. (2017). *On Formalizing Fairness in Prediction with Machine Learning*. Retrieved from http://arxiv.org/abs/1710.03184

Gatys, L., Ecker, A., & Bethge, M. (2016). A Neural Algorithm of Artistic Style. *Journal of Vision*, *16*(12), 326. https://doi.org/10.1167/16.12.326

Glymour, B., & Herington, J. (2019). Measuring the biases that matter the ethical and casual foundations for measures of fairness in algorithms. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 269–278. https://doi.org/10.1145/3287560.3287573

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Retrieved from http://www.deeplearningbook.org/

Green, B., & Hu, L. (2018). The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. *The ICML 2018 Debates Workshop*. Retrieved from https://econcs.seas.harvard.edu/files/econcs/files/green_icml18.pdf

Greenberg, I. (1979). An Analysis of the EEOCC "Four-Fifths" Rule. *Management Science*, *25*(8), 762–769. https://doi.org/10.1287/mnsc.25.8.762

Guszcza, J., Rahwan, I., Bible, W., Cebrian, M., & Katyal, V. (2018). Why We Need to Audit Algorithms. Retrieved July 14, 2020, from Harvard Business Review website: https://hbr.org/2018/11/why-we-need-to-audit-algorithms

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*. Retrieved from http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Hannák, A., Mislove, A., Wagner, C., Strohmaier, M., Garcia, D., & Wilson, C. (2017). Bias in Online freelance marketplaces: Evidence from TaskRabbit and Fiverr. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 1914–1933. https://doi.org/10.1145/2998181.2998327

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in*

Neural Information Processing Systems, 3323–3331. Neural information processing systems foundation.

Hu, L., & Chen, Y. (2018, November 30). *A Short-term Intervention for Long-term Fairness in the Labor Market*. 1389–1398. https://doi.org/10.1145/3178876.3186044

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2016). *Better Fair Algorithms for Contextual Bandits*. Retrieved from https://arxiv.org/abs/1610.09559.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kamiran, F., Karim, A., & Zhang, X. (2012). *Decision Theory for Discrimination-aware Classification*. https://doi.org/10.1109/ICDM.2012.45

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7524 LNAI*(PART 2), 35–50. https://doi.org/10.1007/978-3-642-33486-3_3

Karras, T., Laine, S., & Aila, T. (2018). *A Style-Based Generator Architecture for Generative Adversarial Networks*. Retrieved from http://arxiv.org/abs/1812.04948

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *35th International Conference on Machine Learning, ICML 2018*, *6*, 4008–4016. International Machine Learning Society (IMLS).

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, *2017-Decem*, 657–667. Retrieved from http://arxiv.org/abs/1706.02744

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Leibniz International Proceedings in Informatics, LIPIcs*, *67*. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Koulish, R. (2016). Immigration Detention in the Risk Classification Assessment Era. *Connecticut Public Interest Law Journal*, *16*(1). Retrieved from https://www.congress.gov/bill/114th-

congress/house-bill/2029/text.

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). *Counterfactual Fairness*. Retrieved from http://arxiv.org/abs/1703.06856

Kusner, M. J., Russell, C., Loftus, J. R., & Silva, R. (2018). *Causal Interventions for Fairness*. Retrieved from http://arxiv.org/abs/1806.02380

Lange, J. E., Johnson, M. B., & Voas, R. B. (2005). Testing the racial profiling hypothesis for seemingly disparate traffic stops on the New Jersey Turnpike. *Justice Quarterly*, *22*(2), 193–223. https://doi.org/10.1080/07418820500088952

Lansing, S. (2012). New York State COMPAS-Probration Risk and Need Assessment Study: Examining the Recidivism Scale's Effectiveness and Predictive Accuracy. In *Division of Criminal Justice Services Office of Justice Research and Performance*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. In *ProPublica*. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Liu, S., & Vicente, L. N. (2020). Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *ArXiv*.

Loftus, J. R., Russell, C., Kusner, M. J., & Silva, R. (2018). *Causal Reasoning for Algorithmic Fairness*. Retrieved from http://arxiv.org/abs/1805.05859

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 502–510. https://doi.org/10.1145/2020408.2020488

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 349–358. https://doi.org/10.1145/3287560.3287564

Maguolo, G., & Nanni, L. (2020). *A Critic Evaluation of Methods for COVID-19 Automatic Detection from X-Ray Images*. Retrieved from http://arxiv.org/abs/2004.12823

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). *A Survey on Bias and Fairness in Machine Learning*. Retrieved from http://arxiv.org/abs/1908.09635

Miconi, T. (2017). *The impossibility of "fairness": a generalized impossibility result for decisions*. Retrieved from http://arxiv.org/abs/1707.01195

Naudts, L. (2018). Towards Accountability: The Articulation and Formalization of Fairness in Machine Learning. In *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3298847

Ng, A. (2018). *Machine Learning Yearning*. Retrieved from deeplearning.ai

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). *Deep Learning for Deepfakes Creation and Detection*. Retrieved from http://arxiv.org/abs/1909.11573

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., … Edwards, B. (2018). Adversarial Robustness Toolbox v1.0.0. *ArXiv*. Retrieved from http://arxiv.org/abs/1807.01069

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00013

Oneto, L., Donini, M., Maurer, A., & Pontil, M. (2019). *Learning Fair and Transferable Representations*. Retrieved from http://arxiv.org/abs/1906.10673

Pager, D., & Shepherd, H. (2008). The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. *Annual Review of Sociology*, *34*(1), 181–209. https://doi.org/10.1146/annurev.soc.33.040406.131740

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware Data Mining. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 560–568. https://doi.org/10.1145/1401890.1401959

Pessach, D., & Shmueli, E. (2020). Algorithmic fairness. *ArXiv*. https://doi.org/10.1257/pandp.20181018

Petrocelli, J. V., & Sherman, S. J. (2010). Event detail and confidence in gambling: The role of counterfactual thought reactions. *Journal of Experimental Social Psychology*, *46*(1), 61–72. https://doi.org/10.1016/j.jesp.2009.09.013

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On Fairness and Calibration. *Advances in Neural Information Processing Systems*, *2017-December*, 5681–5690. Retrieved from http://arxiv.org/abs/1709.02012

Potash, E., Brew, J., Loewi, A., Majumdar, S., Reece, A., Walsh, J., … Ghani, R. (2015). Predictive modeling for public health: Preventing childhood lead poisoning. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2015-Augus*, 2039–2047. https://doi.org/10.1145/2783258.2788629

PyMetrics. (2017). pymetrics/audit-ai: detect demographic differences in the output of machine learning models or other assessments - Github repository. Retrieved December 4, 2019, from https://github.com/pymetrics/audit-ai

Rambachan, A., Kleinberg, J., Ludwig, J., & Mullainathan, S. (2020). *An Economic Approach to Regulating Algorithms **. Retrieved from https://pdfs.semanticscholar.org/1672/73526caeae8439b2c08630f984b758f76dab.pdf

Roese, N. J., & Olson, J. M. (1996). Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration. *Journal of Experimental Social Psychology*. https://doi.org/10.1006/jesp.1996.0010

Rothwell, J. (2014). How the War on Drugs Damages Black Social Mobility. *The Brookings Institution*. Retrieved from https://www.brookings.edu/blog/social-mobility-memos/2014/09/30/how-the-war-on-drugs-damages-black-social-mobility/

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., … Ghani, R. (2018). *Aequitas: A Bias and Fairness Audit Toolkit*. Retrieved from http://arxiv.org/abs/1811.05577

Sapiezynski, P., Wilson, C., & Kassarnig, V. (2017). Academic performance prediction in a gender-imbalanced environment. *Proceedngs of FATRECWorkshop on Responsible Recommendation at*

ACM RecSys, Como, Italy, (August), 15–18. https://doi.org/10.18122/B20Q5R

Schermer, B. W., Hagenauw, D., & Falot, N. (2018). *Handleiding Algemene verordening gegevensbescherming*. Retrieved from https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/handleidingalgemeneverordeninggegevensbescherming.pdf

Shrestha, Y. R., & Yang, Y. (2019). Fairness in Algorithmic Decision-Making: Applications in Multi-Winner Voting, Machine Learning, and Recommender Systems. *Algorithms*, *12*(9), 199. https://doi.org/10.3390/a12090199

Suresh, H., & Guttag, J. V. (2019). *A Framework for Understanding Unintended Consequences of Machine Learning*. Retrieved from http://arxiv.org/abs/1901.10002

Tkachenko, Y. (2015). *Autonomous CRM Control via CLV Approximation with Deep Reinforcement Learning in Discrete and Continuous Action Space*. Retrieved from http://arxiv.org/abs/1504.01840

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., & Jiang, N. (2013). Children in the Public Benefit System at Risk of Maltreatment Identification Via Predictive Modeling. *American Journal of Preventive Medicine*, *45*(3), 354–359. https://doi.org/10.1016/j.amepre.2013.04.022

Verma, S., & Rubin, J. (2018). Fairness Definitions Explained. *IEEE/ACM International Workshop on Software Fairness*, *18*. https://doi.org/10.1145/3194770.3194776

Vijayakumar, S. (2018). *Interpretability Through Interrogation: Fairness and Interpretability in the Context of Criminal Sentencing*. Harvard College.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3063289

Weinberger, D. (2018). Playing with AI Fairness. Retrieved December 4, 2019, from https://pair-code.github.io/what-if-tool/ai-fairness.html

Weymark, J. (2016). *Social Welfare Functions* (Vol. 1; M. D. Adler & M. Fleurbaey, Eds.). https://doi.org/10.1093/oxfordhb/9780199325818.013.5

Winick, E. (2018). This company audits algorithms to see how biased they are - MIT Technology Review. Retrieved November 5, 2019, from https://www.technologyreview.com/f/611113/a-new-company-audits-algorithms-to-see-how-biased-they-are/

Wu, Y., Zhang, L., & Wu, X. (2019). Counterfactual Fairness: Unidentification, Bound and Algorithm. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, 1438–1444. https://doi.org/10.24963/ijcai.2019/199

Wu, Y., Zhang, L., Wu, X., & Tong, H. (2019). *PC-Fairness: A Unified Framework for Measuring Causality-based Fairness*. Retrieved from http://arxiv.org/abs/1910.12586

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *26th International World Wide Web Conference, WWW 2017*, 1171–1180. https://doi.org/10.1145/3038912.3052660

Zemel, R., Ledell, Y. (, Wu,), Swersky, K., Pitassi, T., & Dwork, C. (2013). *Learning Fair Representations*. Retrieved from PMLR website: http://proceedings.mlr.press/v28/zemel13.html

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. Retrieved from http://arxiv.org/abs/1801.07593

Appendices

Appendix A – Descriptive statistics on biased dataset

A.1 Dataset 2: Disparate Impact on FNR

An unconstrained logistic regression classifier was trained on this training data, using only x_1 and x_2 to predict the *outcome_of_interest*. The output of this model in term of probabilities is called *proba* and the resulting binary prediction is called decision. This classifier attains a test set accuracy of 0.81, but leads to a difference in False Negative Rate D_{FNR} = 0.53 - 0.0 = 0.531, which constitutes a clear case of disparate mistreatment in terms of False Negative Rates. D_{FNR} = 0.19 - 0.19 = 0. The social welfare score for this untreated dataset is 0.80.

Variable	Mean	Standard deviation
outcome_of_interest	0.50	0.50
sensitive_att	0.50	0.50
x_1	0.75	2.36
x_2	0.75	2.42
decision	0.46	0.50
proba	0.50	0.31

Table A1: Descriptive statistics for biased dataset 2 (training data).

Variable	Mean	Standard deviation
outcome_of_interest	0.50	0.50

sensitive_att	0.50	0.50
x_1	0.75	2.35
x_2	0.75	2.43
decision	0.46	0.50
proba	0.50	0.31

Table A2: Descriptive statistics for biased dataset 2 (test data).

A.2 Dataset 3: Disparate Impact on both FPR and FNR (different sign)

An unconstrained logistic regression classifier was trained on this training data, using only x_1 and x_2 to predict the *outcome_of_interest*. The output of this model in term of probabilities is called *proba* and the resulting binary prediction is called decision. This classifier attains a test set accuracy of 0.81, but leads to a difference in False Negative Rate D_{FNR} = 0.09 - 0.29 = -0.19 and to a difference in False Positive Rate D_{FPR} = 0.29 - 0.10 = 0.20, which constitutes a clear case of disparate mistreatment in terms of both False Positive Rates and False Negative Rates. The social welfare score for this untreated dataset is 0.81.

Variable	**Mean**	**Standard deviation**
outcome_of_interest	0.50	0.50
sensitive_att	0.50	0.50
x_1	0.50	2.68
x_2	-0.01	3.07

decision	0.50	0.50
proba	0.50	0.34

Table A3: Descriptive statistics for biased dataset 3 (training data).

Variable	**Mean**	**Standard deviation**
outcome_of_interest	0.50	0.50
sensitive_att	0.50	0.50
x_1	0.51	2.68
x_2	0.01	3.08
decision	0.50	0.50
proba	0.50	0.34

Table A4: Descriptive statistics for biased dataset 3 (test data).

A.3 Dataset 4: Disparate Impact on both FPR and FNR (same sign)

An unconstrained logistic regression classifier was trained on this training data, using only x_1 and x_2 to predict the *outcome_of_interest*. The output of this model in term of probabilities is called *proba* and the resulting binary prediction is called decision. This classifier attains a test set accuracy of 0.82, but leads to a difference in False Negative Rate D_{FNR} = 0.11 - 0.23 = -0.12 and to a difference in False Positive Rate D_{FPR} = 0.06 - 0.31 = -0.25, which constitutes a clear case of disparate mistreatment in terms of both False Positive Rates and False Negative Rates. The social welfare score for this untreated dataset is 0.82.

Variable	Mean	Standard deviation
outcome_of_interest	0.50	0.50
sensitive_att	0.50	0.50
x_1	-0.50	3.74
x_2	1.00	2.78
decision	0.51	0.50
proba	0.50	0.35

Table A5: Descriptive statistics for biased dataset 4 (training data).

Variable	Mean	Standard deviation
outcome_of_interest	0.50	0.50
sensitive_att	0.50	0.50
x_1	-0.48	3.74
x_2	1.01	2.78
decision	0.51	0.50
proba	0.50	0.35

Table A6: Descriptive statistics for biased dataset 4 (test data).

Appendix B – Tables showing trade-offs between accuracy and social welfare score for each biased dataset

	Dataset 1	
	Accuracy	Social Welfare Score
Disparate Impact Remover	+	+
Reweighing	0	+
Meta Algorithm for Fair Classification	-	-
Prejudice Remover	+	+
Calibrated Equalized Odds	-	-
Reject Option Classification	0	-

Table B1: Evaluating of changes in accuracy and social welfare with respect to the untreated dataset 1

	Dataset 2	
	Accuracy	Social Welfare Score
Disparate Impact Remover	+	+
Reweighing	-	-
Meta Algorithm for Fair Classification		
Prejudice Remover	0	+
Calibrated Equalized Odds	-	-
Reject Option Classification	-	-

Table B2: Evaluating of changes in accuracy and social welfare with respect to the untreated dataset 2

	Dataset 3	
	Accuracy	Social Welfare Score
Disparate Impact Remover	+	+
Reweighing	0	0
Meta Algorithm for Fair Classification	-	-
Prejudice Remover	+	+
Calibrated Equalized Odds	-	-
Reject Option Classification	+	+

Table B3: Evaluating of changes in accuracy and social welfare with respect to the untreated dataset 3

	Dataset 4	
	Accuracy	Social Welfare Score
Disparate Impact Remover	-	-
Reweighing	+	0
Meta Algorithm for Fair Classification	-	-
Prejudice Remover	+	+
Calibrated Equalized Odds	-	-
Reject Option Classification	0	-

Table B4: Evaluating of changes in accuracy and social welfare with respect to the untreated dataset 4