Janssen, Patrick; Sadowski, Bert M.

Conference Paper

Bias in Algorithms: On the trade-off between accuracy and fairness

23rd Biennial Conference of the International Telecommunications Society (ITS): "Digital societies and industrial transformations: Policies, markets, and technologies in a post-Covid world", Online Conference / Gothenburg, Sweden, 21st-23rd June, 2021

Provided in Cooperation with:

International Telecommunications Society (ITS)

Suggested Citation: Janssen, Patrick; Sadowski, Bert M. (2021) : Bias in Algorithms: On the trade-off between accuracy and fairness, 23rd Biennial Conference of the International Telecommunications Society (ITS): "Digital societies and industrial transformations: Policies, markets, and technologies in a post-Covid world", Online Conference / Gothenburg, Sweden, 21st-23rd June, 2021, International Telecommunications Society (ITS), Calgary

This Version is available at:
https://hdl.handle.net/10419/238032

Bias in Algorithms: On the trade-off between accuracy and fairness

Patrick Janssen[1], *Bert M. Sadowski* [2,3,4]

[1] *Rabobank, The Netherlands*

[2] *School of Innovation Sciences,* Eindhoven University of Technology, *The Netherlands*

[3] *Newcastle Business School, Northumbria University, Newcastle, UK*

[4] *Jiangsu University, Zhenjiang, China*

Working document. Please, do not cite.

Paper submitted to the ITS Biennial Conference 2021, 21st – 23rd June, Gothenburg, Sweden.

Abstract

Within the discussion on bias in algorithmic selection, fairness interventions are increasingly becoming a popular means to generate more socially responsible outcomes. The paper uses a modified framework based on Rambachan et. al. (2020) to empirically investigate the extent to which bias mitigation techniques can provide a more socially responsible outcome and prevent bias in algorithms. In using the algorithmic auditing tool AI Fairness 360 on a synthetically biased dataset, the paper applies different bias mitigation techniques at the preprocessing, inprocessing and postprocessing stage of algorithmic selection to account for fairness. The data analysis has been aimed at detecting violations of group fairness definitions in trained classifiers. In contrast to previous research, the empirical analysis focusses on the outcomes produced by decisions and the incentives problems behind fairness.

The paper showed that binary classifiers trained on synthetically generated biased data while treating algorithms with bias mitigation techniques leads to a decrease in both social welfare and predictive accuracy in 43% of the cases tested. The results of our empirical study demonstrated that fairness interventions, which are designed to correct for bias often lead to worse societal outcomes. Based on these results, we propose that algorithmic selection involves a trade-between accuracy of prediction and fairness of outcomes. Furthermore, we suggest that bias mitigation techniques surely have to be included in algorithm selection but they have to be evaluated in the context of welfare economics.

Introduction

As algorithms based on machine learning are increasingly becoming popular in facilitating decision-making processes in a variety of industries (Boodhun & Jayabalan, 2018; Coussement et al., 2017), criticism is mounting with respect to emerging bias in a number of application areas and an urge to better understand algorithmic selection (Awad et al., 2018; O'Neil, 2017). With algorithms are driven by even greater complexity, they are developed with the objective to more precisely predict patterns in larger and large datasets (Tsai et al., 2015; Zhou et al., 2017). Algorithms are utilized to generate more accurate predictions taking a large variety of factors into account, e.g. suggesting whether a candidate is suitable for a job (Raghavan et al., 2020), a debtor will default on his loan (Coussement et al., 2017; Malhotra & Malhotra, 2003), a chest X-ray image show signs of COVID-19 (Maguolo & Nanni, 2020) or whether a Facebook post does not align with community guidelines (Ng, 2018). It has been suggested that fairness interventions can mitigate the negative effects stemming from bias in algorithmic selection (Rambachan et al., 2020; Zafar et al., 2019).

As bias is a common problem in algorithmic selection, remedies of algorithmic bias are seldom explored. As there is a strong belief that greater availability of data sets and better fairness definitions will solve the problem of algorithmic bias, the literature has just recently addressed the issues of fairness and accountability (Rambachan et al., 2020). In contrast to conventional wisdom, we argue that greater availability of data and better fairness definition are a necessary but not sufficient step to facilitate the structure and evolution of algorithms. In using a modified framework based on Rambachan et. al. (2020), we empirically investigate the extent to which it is socially desirable to implement bias mitigation techniques to combat bias in algorithms. To streamline the process of implementing these bias mitigation techniques, the audit tool 30 AI Fairness 360 (AIF360) by IBM is used. By using this algorithmic audit tool, we are able to detect violations of group fairness definitions in trained classifiers. In order to examine the effects of fairness interventions, we applied different bias mitigation techniques at the preprocessing, inprocessing and postprocessing stage to a synthetically biased dataset. The dataset was generated as proposed by Zafar et al. (2017) using a value for *n* of 500.000, allowing for greater accuracy while still being computational feasible.

In the following, we discuss the literature on bias and responsible algorithm. In linking the discussion to the social welfare function as derived from (Rambachan et al., 2020), we show that it becomes possible to measure the societal preference of outcomes of an algorithm (section 2). Furthermore, we develop a framework for examining the different fairness definitions at the different stages of data processing (section 3). Afterwards, we present the results of the empirical study by applying different fairness interventions (section 4). The paper concludes with some policy and managerial implications (section 5).

2. Theoretical discussion: Fairness and algorithmic prediction

2.1. Fairness and social welfare

Within the current literature, there has been a consensus that fairness has to be studied as the endpoint of implementing algorithms (Pleiss et al., 2017). As the implementation of algorithms is rooted in the transformation of observed data into predictions or decisions. These transformations or mappings are then formally investigated as whether or not these mappings are "fair." (Pleiss et al., 2017). From a particular definition of fairness, research has then focused on constructing fair mappings from data and the extent to which the algorithm fits to the definition utilized (Chouldechova, 2017; Dwork et al., 2012; Zemel et al., 2013). However, the literature has convincingly proposed that algorithms require regulation as well as developers/ creators should develop more fair algorithms (Doshi-Velez et al., 2017).

In contrast to this research tradition, fairness has more recently been defined in terms of preferences over the resulting outcomes of the screening decision using a social welfare function (Rambachan et al., 2020). By using a welfare-economics approach on the regulation of algorithms Rambachan et al (2020) define a theoretically optimal algorithmic regulation, in which bias regarding decision-making systems can be reduced in contrast to a world in which all decisions are made solely by humans. They propose that in order to provide for optimal regulation decision-makers have to disclose their predictive algorithm, the data utilized for the training of the algorithm, and the *decision rule*, which they call the *social planner*. Based on the decision rule, the prediction made by an algorithm is defined and it is assigned to a particular decision. A predicted credit score, for example, is assigned a certain threshold above which the loan of a client is approved. A social planner would be the party that is concerned with optimizing social welfare and has control over public policy regulation regarding this algorithm. This can be a government institution responsible for regulating the algorithms used by the firm. The algorithm faces a *regulation problem*; it can be restricted with respect to the attributes the decision-makers will use in the predictions, but regulation has actually no impact on the design and implementation of the actual algorithms.

In their model Rambachan et al (2020) propose, that social welfare is maximised in a situation, when discrimination is zero, as the decision-makers must disclose the decision rules. In this case, algorithms are regulated based on input regulations. However, as it still remains unclear how the relationship between input data and (biased) outcomes is defined, this approach might not be the best solution to the problem of regulation. In this context, Cowgill & Tucker (2019) provide a different solution to the problem by advocating to regulate the output, leaving the implementation up to the decision-makers themselves (Cowgill & Tucker, 2019). As this approach is more concerned with reducing bias rather than complying with a fair process, the implementation is left to the private sector.

However, both approaches are limited in certain respects: Firstly, assuming regulation is in place making algorithmic audits mandatory, fairness audits have to be performed by a certain agency. However, the responsibilities for an agency involved in auditing decisions are not clearly defined yet. Within Europe, further progress has been made with the General Data Protection regulation of 2018 (European Commission, 2018) and the 'right to explanation' embedded within the GDPR as mandating algorithmic audits (Edwards & Veale, 2017). In the recently proposed Artificial

Intelligence Act, the European Commission proposes that newly established national public institutions have to "promote public trust in the use of AI and strengthen enforcement mechanisms (by introducing a European coordination mechanism, providing for appropriate capacities, and facilitating audits of the AI systems with new requirements for documentation, traceability and transparency)." (European Commission, 2021).

Secondly, if algorithmic audit is related to a full disclosure of the data, the trained predictive algorithm and the corresponding decision rules, an optimal social welfare level can be achieved and the bias can be reduced. However, the current auditing techniques for algorithms consist of different kind of specialized software tools which come to a variety of outcomes. These variety of outcomes will even be achieved if the predictive algorithm, training data and/or predictions made by the algorithm are tested against fairness definitions. There are a large number of fairness definitions which are not all compatible with each other. At this stage, different outcomes can be achieved which might not even guarantee that the bias has been eradicated. In this respect, certain predictive algorithm which are considered as fair by one definition may still contain a bias according to another fairness definition.

2.2. Trade-off between fairness and prediction

In order to account for the shortcomings of current outcomes with respect to fairness, researchers have proposed that there actually is a trade-off between achieving different fairness measures and accuracy in predictive modelling. This trade-off has been due to the fact that different methods for fair machine learning place additional constraints on algorithms or penalize certain ways of algorithms learn (Berk et al., 2017; Liu & Vicente, 2020). This also means that this trade-off has to be placed in the context of multiple definitions of fairness and fairness interventions. The basic assumption in this literature is that making algorithms fairer comes at the cost of predictive accuracy.

Withing the economics discipline, this trade-off has to be placed at the center of a social welfare function in order to compare social alternatives. Based on a social welfare function, the preferences of each individual in a society can be aggregated, which allows to provide guidance with respect to decisions that influence social welfare. Rambachan et al. (2020) apply this framework to propose an approach to regulate fairness in algorithms. Based on their social welfare definition, fairness is related to the preferences of society with respect to the outcomes of the decision-making process. The optimal outcome for society in this regulation process is a function in terms of the *outcome of interest* used in creating the particular algorithm. They define the social welfare function as the weighted average outcome of interest among individuals that receive a 'positive' decision by the predictive algorithm. That means, it can be optimal when the total predicted productivity of a hired candidate is maximized. The social welfare function contains the *weighted* average of the different 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 should be increased. As hiring a female candidate would increase social welfare compared to hiring a male candidate (assuming the rest of their characteristics are exactly identical). In this case, the socially optimal decision would be 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. The question still remains however, whether or not this decision reduces bias at the cost of better prediction.

2.3. Bias in algorithmic prediction

The extensive literature on algorithmic bias has rarely included issues of fairness in algorithmic selection (Doshi-Velez & Kim, 2017). However, this literature has demonstrated that the application of algorithms in decision-making processes is often leading to bias (Mehrabi et al., 2019; Olteanu et al., 2019). The bias in data analysis can arise at different phases in the process of applying algorithms: in the data generation phase, in the model building phase or in the implementation phase. Suresh and Guttag (2019) make a distinction between different types of biases and the way they can influence the decisions made by an algorithm (Suresh & Guttag, 2019). According to their categorization, different types of bias can be described as:

- *Historical bias*: The assumption is that even in cases where 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 different kind of outcomes. As the real world still reflects historical issues such as prejudice and stereotyping, they will also be included in the data analysis.

- *Representation bias*: This bias can is related to *selection bias*, in which a particular distribution of a sample (on which the model is trained) actually does not match the real distribution of the population in the real world. This can be due to insufficient sampling methods, which do not include all groups, or in cases, when the population of interest does not match the training data set.

- *Measurement bias*: This bias can be found when certain features and labels utilized in a model are just *proxies* for actual features and labels in the real world. As these proxies might have different distributions for different groups, this can lead to different biases across groups

- *Aggregation bias*: This form of bias occurs when a particular model is used to describe multiple groups, which actually have quite different conditional distributions. This can lead to a situation where the mapping from inputs to outputs is rather different for other groups.

- *Evaluation bias*: This bias is rooted in the need to objectively compare different models to each other. In order to compare these models, standardized benchmarks are required. This allows to optimize models with respect to their training data and the external benchmarks used (Suresh & Guttag, 2019).

2.4. Fairness in algorithmic prediction: Group fairness vs. individual fairness

Within the literature on algorithm prediction, the scientific discussion has not focused on bias, but more recently on the extent to which algorithms create *fair* outcomes for the individuals concerned. Despite a growing research in the area, there currently is not a generally accepted definition of a fair algorithm as the debate has concentrated on formalizing various notions of fairness. In more general terms, the notions of fairness can be classified in two distinct categories: group fairness and individual fairness (Gajane & Pechenizkiy, 2017; Naudts, 2018; Verma & Rubin, 2018).

In order to achieve group fairness, different groups of individuals should be treated equally (Dwork et al., 2012; Kusner et al., 2017). There are different notions of group fairness with respect to the exact metric that the groups should be compared to, including group-independent predictions

(Pedreshi et al., 2008), statistical parity (Pager & Shepherd, 2008), equality of opportunity (Hardt et al., 2016) or predictive rate parity(Verma & Rubin, 2018). Individual fairness, in contrast, refers to an similar treatment of individuals who exhibit similar characteristics. In this case, the algorithm should predict the same outcome. Individual fairness requires 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). Still it remains unclear, however, how to determine such a distance metrics, which measures the similarity between two inputs (individuals).

Research has shown that is impossible to satisfy the three most prevalent notions of group fairness at the same time. These three notions are to equalize the odds, achieve statistical parity and generate predictive rate parity (Chouldechova, 2017; Miconi, Thomas, 2017; Pleiss et al., 2017). Based on this research, it becomes important to focus on the trade-off between statistical parity, equality of opportunity and predictive rate parity. In reality, this demonstrates that an algorithm will always involve some type of bias, as fulfilling the requirements of one of these three definitions means violating at least one of the other two (Dieterich et al., 2016; Lansing, 2012; Larson et al., 2016).

3. Conceptual framework and methodology: Fairness solutions to the problem of bias in algorithms

3.1 European Regulation on auditing and detecting bias in algorithms

Within Europe, the General Data Protection Regulation (GDPR) of 2018 provides the legal basis for detecting bias in algorithms and auditin (European Commission, 2018). Under the GDPR, personal data can only be processed if at least 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. As it has been suggested that the GDPR also provides a right to explanation, i.e. enables citizens to get more information on workings of the algorithms, which influence their lives, the work and how they make their predictions. However, currently it is still unclear how this right should be implemented and enforced (Casey et al., 2019; Wachter et al., 2017). In the Artificial Intelligence Act (European Commission, 2021), the European Commission went a step further and defined fundamental rights of EU citizens based on transparency, accountability, non-discrimination and auditability of algorithms.

3.2 Auditing and detecting bias in algorithms

In case a bias in an algorithm has been detected by an audit, the next step is removing it. Measures that enhance fairness by removing bias are different fairness(-enhancing) interventions or bias mitigation techniques. Researchers generally make a distinction between the methods used in removing bias based on where in the modelling pipeline they are used. In *pre-processing* 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).

Pre-processing methods: Pre-processing 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 pre-processing method (Feldman et al., 2014). A promising variant of pre-processing 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).

3.3 Examining algorithm fairness: Hypotheses

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. 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.

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. The following hypothesis will be tested:

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

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.

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

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.

3.4 Creating the biased synthetic dataset

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.

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.

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)

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.

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.

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.*

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).

3.5 Training classifiers on datasets

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.

3.6 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.

In the paper, the procedure to analyse the extent to which bias mitigation techniques have an impact on social welfare were as follows:

1. Generate a 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

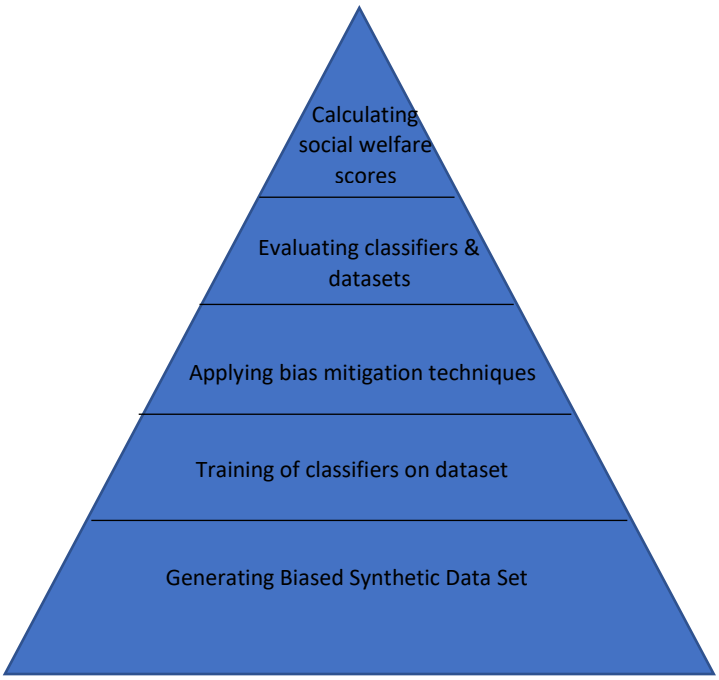

Figure 1: Procedure to apply bias mitigation techniques to analyze social welfare improvements

To undertake the empirical analysis, the algorithmic audit tool AIF360 was used to implement bias mitigation techniques. Table 2 shows the bias mitigation techniques supported by AIF360 and whether (or not) they are used in this research. In contrast to the discussion on auditing predicted risk scores for defendants (Dieterich et al., 2016)(Green & Hu, 2018), the methodology allows to directly measure the social impact through the use of a social welfare function.

Type of intervention	Bias mitigation technique	Used in this research
Pre-processing	Disparate Impact Remover (Feldman et al., 2015)	Yes
	Reweighing (Kamiran & Calders, 2012)	Yes
	Learning Fair Representations (Zemel et al., 2013)	No
	Optimized Preprocessing (Calmon et al., 2017)	No
Inprocessing	Prejudice Remover (Kamiran & Calders, 2012)	Yes
	Meta Algorithm for Fair Classification (Celis et al., 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 & Calders, 2012)	Yes

Table 2: ***Type of intervention and bias mitigation technique***

To evaluate the different bias mitigation techniques with respect to their social outcomes, social welfare scores were calculated for a number of biased datasets before and after bias mitigation techniques were applied. The dataset can be considered as a pool of candidates for a screening decision, with every dataset being biased in a different way. The dataset included in each row a single candidate with certain attributes (like *outcome_of_interest*, *sensitive_att*, x_1 and x_1). Within the dataset, there were 500 000 candidates, with half of candidates were part of sensitive attribute group 0, and the other half belong to a sensitive attribute group 1. Furthermore, the datasets were split into a training set and a test set with a ratio of 80/20.

Pre-processing methods

Disparate Impact Remover

Disparate Impact Remover was introduced by Feldman et al. (2014) and is a pre-processing 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.

Reweighing

Reweighing is a pre-processing 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.

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.

Optimized Pre-processing

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.

Inprocesssing methods

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).

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.

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.

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.

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.

Postprocessing methods

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 and changes the outputted predictions to achieve equalized odds.

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.

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:

$$ACC = TP + TNTP + TN + FP + FN$$

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.

4. Empirical Analysis: Fairness Definitions and Welfare Scores

4.1 Calculating social welfare scores before intervention

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.

In a first step, an unconstrained logistic regression classifier was trained on a training data set, by just using x_1 and x_2 to predict the attribute OUT (*outcome_of_interest*). The output of the regression model in term of probabilities (called PROBA) and the resulting binary prediction is called decision. This classifier attained a test set accuracy of 0.77. But it actually increased the 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
x1	0.75	2.36
x2	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
x1	0.76	2.36
x2	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 training and the test set are 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**	
Pre-processing	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 impacton 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**	
Pre-processing	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).

Pre-processing 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.

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.

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.

4.2 Calculating social welfare scores after intervention

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.

Intervention		Social welfare score	Change in social welfare score
Before intervention		**0.75**	
Pre-processing	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%
	Calibrated Equalized Odds	0.60	-19.6%

Postprocessing	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).

4.3 Disparate mistreatment

As discussed earlier, 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**	***DACC***	**Change in *DACC***
Before intervention		**0.81**	**0.64**	**0.90**	**0.26**	
Pre-processing	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 exhibitsdisparate impact on FNR (test data).

Pre-processing 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.

Inprocessing methods

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.

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.

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.

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). 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**	
Pre-processing	Disparate Impact Remover	**0.81**	**0.81**	**0.81**	**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**	
Pre-processing	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).

Pre-processing 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.

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}. Prejudice Remover has almost equal results as Disparate Impact Remover.

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.

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)

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.

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). 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.

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)

Pre-processing 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. Reweighing does not yield any improvements for this dataset.

Inprocessing methods

The 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.

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}.

Social Welfare Scores

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.

4.4 Trade-off between accuracy and change in social welfare

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 WelfareScore	**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 Scoresincrease, 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.

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.

5 Summary and conclusions

The objective of the paper was to examine whether or not the application of bias mitigation techniques will lead to more societally desirable outcomes. Using predictive algorithms to aid in screening decision-making opens this process up to algorithmic bias, which leads to a sub-optimal outcomes. This is both socially and economically undesirable; decision-makers gain higher benefits from optimal screening decisions. This bias can arise in multiple stages in the modelling process; 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.

In line with Rambachan et al. (2020) the paper used a welfare-economics approach to examine algorithmic bias. By using insights on different types of bias, the paper described how bias can be detected and corrected in algorithms. A social welfare function was used to quantify society's preferences over the outcomes of an algorithmic decision-making process. By using a biased synthetic datasets, different method for analysing bias mitigation techniques through the lens of social welfare functions were utilized. The results of the analysis demonstrated that there is a trade-off between social welfare and predictive accuracy in the context of algorithmic decision-making.

As the analysis was based on a synthetically created biased datasets, bias mitigation techniques were used to analyse this bias applying a social welfare function. By examining the social welfare scores before and after treating each biased model/dataset with the chosen fairness intervention, it was shown whether or not this treatment resulted in more desirable outcomes. This, in turn, creates insights for the regulation of algorithms.

The results from the analysis 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. The results of the empirical analysis show that there is 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).

This research contributed to existing research in different ways. Firstly, this research added an empirical analysis to the welfare economics framework on bias correcting in algorithms. Secondly, as the different fairness interventions were embedded in the context of social welfare, their effects could be measured. Thirdly, by using the method for creating biased synthetic biased datasets from Zafar et al. (2017), an empirical test of a number of different bias mitigation techniques was possible. 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.

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 to different degree. While they might not completely make algorithms free of bias, they are often able at least to make sure some metrics are equal for all groups, so that the algorithms make the same mistakes for all groups.

In addition, the paper showed that the AIF360 tool is very accessible for practitioners. 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 demonstrated that state-of-the-art algorithmic audit tools are already able to remove 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, it can be concluded 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), which is in line with the requirements of the Artificial Intelligence Act of 2021 by the European Commission providing some guidelines of the enforcement powers of a regulatory agency on the national level.

References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2017). A convex framework for fair regression. *ArXiv Preprint ArXiv:1706.02409*.

Boodhun, N., & Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems*, *4*(2), 145–154.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3995–4004.

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 Tech. LJ*, *34*, 143.

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

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

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27–36.

Cowgill, B., & Tucker, C. E. (2019). Economics, fairness and algorithmic bias. *Preparation for: Journal of Economic Perspectives*.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, *7*(7.4), 1.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ArXiv Preprint ArXiv:1702.08608*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, *16*, 18.

European Commission. (2018). *General Data Protection Regulation (L119/1)*. European Commission.

European Commission. (2021). *Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (ARTIFICIAL INTELLIGENCE ACT) and Amending certain Union Legislative Acts SEC 2021 167 final*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). *Certifying and removing disparate impact*. 259–268.

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

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *ArXiv Preprint ArXiv:1610.02413*.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *International Conference on Machine Learning*, 2564–2572.

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. *ArXiv Preprint ArXiv:1703.06856*.

Lansing, S. (2012). New York State COMPAS-probation risk and need assessment study: Examining the recidivism scale's effectiveness and predictive accuracy. *Retrieved March*, *1*, 2013.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, *9*(1).

Liu, S., & Vicente, L. N. (2020). Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *ArXiv Preprint ArXiv:2008.01132*.

Maguolo, G., & Nanni, L. (2020). A critic evaluation of methods for covid-19 automatic detection from x-ray images. *ArXiv Preprint ArXiv:2004.12823*.

Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. *Omega*, *31*(2), 83–96.

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

Miconi, Thomas. (2017). The impossibility of "fairness": A generalized impossibility result for decisions. *1707.01195*.

Naudts, L. (2018). Towards Accountability: The Articulation and Formalization of Fairness in Machine Learning. *IFIP Summer School on Privacy and Identity Management" Fairness, Accountability and Transparency in the Age of Big Data"(20-24 August 2018)(Submitted for Pre-Proceedings)*.

Ng, C. W. (2018). Critical multimodal discourse analyses of news discourse on Facebook and YouTube. *Journal of Asia TEFL*, *15*(4), 1174.

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., & Ludwig, H. (2018). Adversarial Robustness Toolbox v1. 0.0. *ArXiv Preprint ArXiv:1807.01069*.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*, 13.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annu. Rev. Sociol*, *34*, 181–209.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). *Discrimination-aware data mining*. 560–568.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *ArXiv Preprint ArXiv:1709.02012*.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481.

Rambachan, A., Kleinberg, J., Ludwig, J., & Mullainathan, S. (2020). *An Economic Perspective on Algorithmic Fairness*. *110*, 91–95.

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *ArXiv Preprint ArXiv:1901.10002*.

Tsai, C.-W., Lai, C.-F., Chao, H.-C., & Vasilakos, A. V. (2015). Big data analytics: A survey. *Journal of Big Data*, *2*(1), 1–32.

Verma, S., & Rubin, J. (2018). *Fairness definitions explained*. 1–7.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness Constraints: A Flexible Approach for Fair Classification. *J. Mach. Learn. Res.*, *20*(75), 1–42.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). *Learning fair representations*. 325–333.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350–361.